ESEIAAT

Trabajo de Final de Grado

ESTUDIO DE ALGORITMOS DE MONITORIZACIÓN INTELIGENTES PARA APLICACIONES INDUSTRIALES

Grado: Ingeniería de Sistemas Audiovisuales

Data de entrega: 10-06-2019

Estudiante: Albert Yanguas Rovira

Director: Juan Antonio Ortega Redondo

Codirector: Jesús Adolfo Cariño Corrales

Índice

Índice de ilustración

Índice de tablas

1. INTRODUCCIÓN

1.1. OBJETO

Estudio y proposición de algoritmos basados en técnicas de procesado avanzado de señal e inteligencia artificial con el fin de caracterizar y reconocer patrones de malfuncionamiento en un tren de laminación de alambrón. Visualización a tiempo real de los resultados para permitir una minimización de la merma de producción.

1.2. ALCANCE

Durante el estudio se realizarán las siguientes tareas:

- Estudio de técnicas de procesado para caracterizar y reconocer patrones.
- Tratamiento estadístico de los datos a estudiar para la obtención de patrones.
- Monitorización de los patrones obtenidos y visualizar dónde se encuentran los fallos.
- Análisis de los resultados obtenidos mediante el tratamiento estadístico y la posterior visualización. Comentar como se pueden evitar los fallos, evitando el trabajo en ciertos patrones de comportamiento.
- Visualización para la identificación de las anomalías a tiempo real.

1.3. REQUERIMIENTOS

Los requisitos a los cuales está sujeto el estudio son:

- Los datos que se van a estudiar pertenecen a una empresa importante del sector metalúrgico español.
- Se va a tener en cuenta la protección de los datos de la empresa, cambiando los nombres de variables que se van a utilizar para representar los datos.

1.4. JUSTIFICACIÓN Y UTILIDAD

1.4.1. Introducción

El uso ascendente de redes y dispositivos electrónicos y la digitalización de los procesos de producción hacen que las actividades económicas y sociales propagan diariamente grandes cantidades de datos. Según algunas estimaciones, la cantidad de datos producidos en todo el mundo se duplica cada dos años, se espera que aumente de 4,4 zettabytes[1] en 2013 a 44 zettabytes en 2020. Esta gran cantidad de datos o *Big Data* pueden obtenerse de interacciones en la web, transacciones comerciales en línea, redes sociales, registros de telefonía móvil, aplicaciones móviles y sensores en objetos vinculados al IoT (*Internet of Things*)[2] [1].

Al mismo tiempo, las tecnologías de la información y la comunicación (TIC) han ido evolucionando, sobre todo en términos de reducción de los costes de almacenamiento, aumento de la capacidad de red, mejora de las herramientas analíticas y disponibilidad de una informática a la carta de alto rendimiento a través de la nube. Estos avances han permitido almacenar, transmitir y procesar grandes cantidades de datos de forma más económica, rápida y eficaz que antes.

Los datos en estos grandes almacenes de datos, ya sea por sí solos o en combinación con datos recopilados de otras fuentes, pueden ser procesados para identificar patrones y extraer relaciones significativas. Los avances en las técnicas analíticas significan que ahora se puede analizar un número cada vez mayor de tipos de datos, incluso datos no estructurados como textos o vídeos en lenguaje natural. Los conocimientos adquiridos pueden utilizarse para diseñar nuevos productos y servicios, mejorar los procesos de producción, optimizar el marketing o la publicidad o mejorar la toma de decisiones.

Según un estudio publicado por el Parlamento Europeo [1], en los próximos años, cuando el mundo se vuelva cada vez más digital y la cantidad de datos aumente, habrá más oportunidades para explotar los datos. El porcentaje de datos que serían útiles para el análisis crecerá del 22% a más del 35%. El estudio también indica que se ha pronosticado un aumentó los ingresos mundiales de más del 50% por el *Big Data* y *Business Analytics*, de casi 122.000 millones de dólares a 187.000 millones. Las industrias que se verán más beneficiadas por estas tecnologías incluyen la industria manufacturera, la banca y los seguros, el gobierno, los servicios

[1] 1 zettabyte equivale a 1 billón de gigabytes
[2] Del inglés internet de las cosas

profesionales, las telecomunicaciones, la salud, el transporte y la venta al por menor. El mercado de datos en Europa existe y representaba, en 2015, casi 55.000 millones de euros, y ha aumentado a un ritmo del 7% anual. Cerca de un 70% de este mercado de datos se concentra en cinco estados miembros: España, Alemania, Francia, Italia y el Reino Unido [1].

Con toda esta tecnología en auge, la industria manufacturera ha evolucionado a lo que llaman la Industria 4.0. Este nuevo concepto, que está marcado por la robótica, la analítica, la inteligencia artificial, las tecnologías cognitivas, la nanotecnología y el IoT, proporciona acceso a tiempo real a los datos y a la inteligencia de negocio. La integración digital de la información desde distintas fuentes y localizaciones permite llevar a cabo negocios en un ciclo continuo. A lo largo de este ciclo, el acceso a tiempo real está impulsado por el continuo y cíclico flujo de información y acciones entre los mundos físicos y digitales. Este flujo se tiene lugar a través de una serie de pasos iterativos conocidos como PDP[3]. Estos pasos son los siguientes, del mundo físico al digital, se captura la información del mundo físico y se crea un registro de la misma, de digital a digital, la información se comparte y se interpreta utilizando analítica avanzada, análisis de escenarios e inteligencia artificial para descubrir información relevante, y por último, del mundo digital al físico, se aplican algoritmos para traducir las decisiones del mundo digital a datos efectivos, estimulando acciones y cambios en el mundo físico. Este estudio se centrará en el segundo paso del ciclo PDP, de digital a digital, interpretando los datos, recogidos del mundo físico, utilizando técnicas de analítica avanzada para descubrir e identificar patrones de comportamiento y monitorizar estos comportamientos, para asegurar un correcto funcionamiento de los procesos [2] [3].

Cómo beneficia toda esta revolución de datos e industrial a las empresas manufactureras. Anticipándose a fallos, mejorando la calidad del producto, aumentando el rendimiento del proceso, etc.

1.4.2. Proyecto

Antes de empezar cabe remarcar que este estudio se basa en los datos reales de una empresa del sector metalúrgico español, con lo cual, para preservar la privacidad de la empresa, se omitirá toda la información considerada confidencial para poder cumplir con la legislación vigente en materia de protección de datos. Por ello, no se va a entrar en el detalle de las

[3] De las siglas en inglés *Physical-to-Digital-to-Physical*

secciones que forman parte del proceso industrial que se va a analizar, así como tampoco se describirá en profundidad a qué hacen referencia cada una de las variables de las señales que se van a analizar.

Es por todo lo visto en el apartado anterior que desde el centro de investigación MCIA UPC[4], centrados en la transferencia de tecnología de ingeniería eléctrica y electrónica para la industria, junto a IThinkUPC[5] han querido apostar por el mundo de los datos y la Industria 4.0. Actualmente están trabajando en diversos proyectos con distintas empresas industriales. Este estudio, como ya se ha dicho anteriormente, está basado en datos reales procedentes de uno de estos proyectos. Una de estas empresas industriales invito a participar, a MCIA UPC y IThinkUPC, en una iniciativa de innovación que busca determinar la probabilidad de aparición de problemas de manufactura y de conocer las causas que las provocan en una de sus líneas de producción, con el fin de poder revertir las condiciones de afectación negativa al rendimiento del proceso.

El proyecto tiene como objetivo desarrollar una herramienta de análisis, fundamentada en modelos basados en datos, que permita determinar la probabilidad de aparición de problemas de producción, así como la afectación a los mismos de las variables del proceso industrial.

1.4.3. Planta experimental

Para entender los datos que se analizarán se procede a explicar cómo es físicamente la línea de producción y como se divide. El proceso industrial que vamos a analizar es un tren de laminación, más concretamente de alambrón. El proceso de laminación es un método de conformado a través del cual el acero es sometido a temperaturas muy elevadas, alrededor de 1200ºC, en un horno durante un período de tiempo, para posteriormente sufrir una serie de reducciones sucesivas, a causa de pasar por unas cajas compuestas por cilindros de laminación, que acaban formando un diseño específico.

[4] Motion Control and Industrial Applications
[5] Empresa de consultoría y servicios avanzados de software de la UPC

Ilustración 1. Tren de alambrón de Russula en Fortaleza, Brasil [13]

El tren de laminación de alambrón comúnmente se divide en distintas secciones. Estas son el horno de recalentamiento, el tren de desbaste, el tren intermedio, el tren acabador, la formación y el enfriamiento [4] [5].

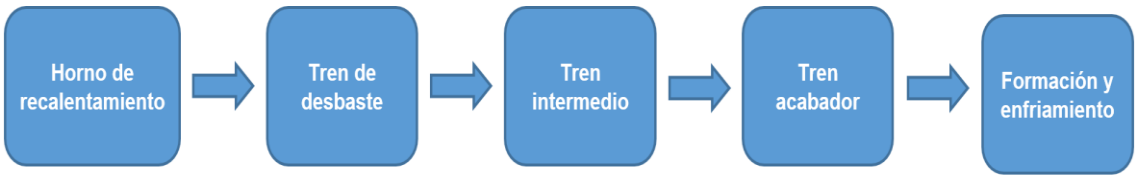

Ilustración 2. Diagrama de un tren de laminación

Para la realización de este estudio vamos a escoger las cinco secciones para el proceso, descritas a continuación, dónde cada una de ellas tiene un conjunto de señales distinto, estas señales son registros sobre el tiempo de distintos sensores conectados a cada una de las secciones respectivas. Estos registros pueden ser la temperatura, presión, posicionamiento, corriente, velocidad, entre otros.

1.4.3.1. Horno de recalentamiento

Tiene una barra empujadora que introduce la palanquilla fría por una ventana lateral. Esta palanquilla empuja a las demás y una ya caliente es forzada a salir por otra ventana situada en la pared opuesta. La palanquilla caliente cae a un transportador de rodillos comandados por el que llega, previamente despuntada, al tren de laminación [6].

Algunas de las variables que se pueden registrar en esta sección son la velocidad de la barra empujadora o la temperatura inicial y final, entre otras.

1.4.3.2. Tren de desbaste

Una vez la palanquilla está a la temperatura adecuada, en el tren de desbastar se realiza un trabajo preparatorio transformando la palanquilla en barras de variadas secciones, con aristas redondeadas. Esta primera operación, además de preparar el lingote para que pueda entrar en el primer perfilador, sirve para homogeneizar bien el metal y para soldar las eventuales sopladuras internas que se hayan podido producir durante la solidificación en la lingotera. Los calibres desbastadores pueden ser de forma rectangular, ojival u ovalada [6].

1.4.3.3. Tren intermedio

Pasando el tren de desbaste, la barra sigue cambiando por el paso del tren intermedio. Este tren se divide en dos, el primer bloque está formado por cuatro cajas como las de un tren de desbaste. El segundo lo forman las “cajas Reynolds” que se caracterizan porque tienen rodillos alternativamente horizontales y verticales. También hay cizallas, carros de bucles, controles, etc. Después del segundo tren intermedio y antes del bloque acabador se tiene el carro de bucles vertical, cizalla, y troceadora [6].

1.4.3.4. Tren acabador

Para darle forma de alambrón, el tren acabador está compuesto por unas diez cajas de laminación. Las cajas son alternativamente óvalo-redondo. Están montadas a 45º con relación a la horizontal, alternativamente arriba y abajo, lo que hace que estén dispuestas en ángulos de 90º entre sí [6].

1.4.3.5. Formación y enfriamiento

El alambrón sale del bloque acabador a unos 1000º C, y se enfría después. Primero es con agua a presión por toberas y después por aire, mientras se desplaza el alambre hacia el formador de rollos Una vez formadas las espiras, el transportador lleva las barras de alambrón hasta las bobinadoras. Hay unos soplantes que efectúan el enfriamiento por aire [6].

2. DESARROLLO DE LA METODOLOGÍA

Para monitorizar el paso de la barra durante el tren de laminado, se propone la siguiente metodología que nos permitirá por medio de técnicas de inteligencia artificial y analítica avanzada identificar desviaciones durante el proceso.

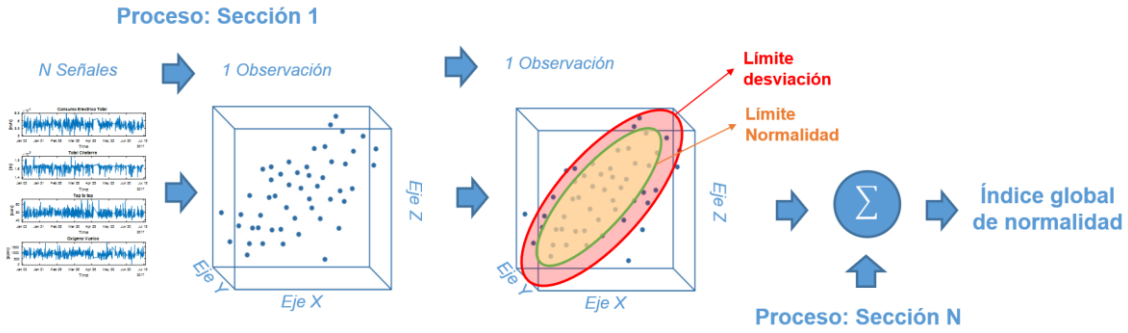

Ilustración 3. Diagrama de la metodología del estudio

La metodología está compuesta de 3 partes: preprocesado, procesado para caracterizar patrones y modelado. El preprocesado consiste en el tratamiento de los datos disponibles del proceso de laminación. El procesado para caracterizar patrones consiste en el uso de técnicas de fusión de información para su análisis. El modelado consiste en la identificación de patrones que nos permitan detectar anomalías o fallos en el proceso de laminación.

2.1. Preprocesado

Este bloque se divide en dos subapartados, el primero, dónde se va a realizar una primera auditoría de los datos utilizados para el análisis. El segundo, será una limpieza de estos datos para un correcto funcionamiento de los algoritmos estadísticos que se utilizan para caracterizar patrones en las señales.

2.1.1. Exploración de datos en crudo

En este bloque se va a desarrollar un estudio exhaustivo de los datos en crudo que se usarán para realizar este estudio. Se van a describir los datos por secciones y se representaran cada una de las señales que se utilizaran para el modelo escogido. Los datos utilizados están formados por dos matrices que se llaman matriz de información y matriz de indicadores, de esta última vamos a tener una por cada sección.

En primer lugar, respecto a la matriz de información, esta matriz se compone de 5.587 observaciones, haciendo referencia al total de palanquillas que se procesaron por la línea de producción durante un periodo de 5 meses. Está formada por 69 columnas, las cuales nos indican la ID del elemento procesado, la hora exacta en que la palanquilla entra al proceso, la línea por la cual pasó la palanquilla, el tipo de fallo, si es que hubo alguno, y otra información que no es relevante para este estudio.

En segundo lugar, haciendo referencia a la matriz de indicadores, contiene información de las señales registradas por cada una de las palanquillas que han sido procesadas. La información que obtenemos de cada señal será el valor medio y el número de muestras. El número de columnas de la matriz dependerá de cada una de las secciones puede ir de 5 a 16, cada una de estas columnas hace referencia a cada sensor que dispone la sección. Por cada señal tenemos el valor medio de la señal durante el paso de la palanquilla, la longitud o número de muestras de la señal durante el paso de la palanquilla.

En este bloque se procede al análisis por sección de cada una de las señales obtenidas, media y longitud, por cada sensor con la intención de poder ver si hay fallos en los registros, causados por el propio sensor o un error informático, o valores que estén muy fuera de la normalidad de la señal que podría afectar a un correcto funcionamiento del modelo. Una manera muy fácil de poder ver estos errores es fijándose en la longitud registrada de la señal, si la longitud registrada para ese elemento y ese sensor es muy diferente a los demás registros, eso significa, que no se ha registrado bien o la muestra no ha llegado a pasar por ese sensor, en el caso que fuera 0. Por último, la media de la señal registrada también es un indicador de errores, fijándonos en la señal física analógica no puede superar ciertos valores, por ejemplo, una temperatura de un horno en funcionamiento deberá estar entre unos valores razonables, una media de 30° será un mal registro.

2.1.2. Limpieza de datos

La técnica de caracterización que se utilizara una vez estén limpios los datos, explicada en el apartado 2.2.1., se basa en la variabilidad de las señales introducidas en el modelo. Es decir, el método es muy sensible a valores anómalos, por lo tanto, es necesario realizar un previo filtrado de datos, con la finalidad de construir un modelo que se ajuste al máximo al modo de trabajo real de las secciones.

Para lograrlo será necesario tener en cuenta el resultado de realizar la exploración de los datos en crudo. Los registros que han sido mal registrados y/o que tienen la longitud o número de muestras igual a cero, son registros que no deberán ser utilizados para el cálculo del modelo, por lo tanto, deberán ser filtrados. Por otro lado, dado que el modelo se basa en la variabilidad de las variables, los *outliers*[6] son muy influyentes en este modelo, con lo cual, también deberán ser filtrados. Las reglas que se van a utilizar para el filtrado de los datos serán las siguientes:

- Limites superiores e inferiores del número de muestras esperado.
- La mitad de las medias de las señales de los diferentes sensores de una sección son igual a cero.
- Dependiendo de los sensores, ajustar unos límites superiores e inferiores de la media de la señal.

2.2. Procesado para caracterizar patrones

Debido a que cada sección tiene 5 señales o más y es complejo realizar una visualización debemos conseguir reducir las dimensiones, pero perdiendo el mínimo de información de las señales. El método llamado *Principal Component Analysis*, permite solventar esta necesidad, fusionando la información y caracterizando el conjunto de las señales en un plano de dos dimensiones.

[6] Observación que es numéricamente distante al resto de los datos, valor atípico.

2.2.1. Principal Component Analysis (PCA)

2.2.1.1. Introducción

Para poder visualizar correctamente una matriz de más de 3 dimensiones y poder buscar ciertos patrones en los datos podemos utilizar la técnica de análisis de componentes principales, PCA[7]. Es un método estadístico que permite simplificar la complejidad de espacios muestrales con muchas dimensiones a la vez que conserva su información. Supóngase que tenemos una muestra con n individuos y p variables $(X_1, X_2, \dots, X_p)$, es decir p dimensiones. El método PCA permite encontrar un número de factores inferior $(z < p)$ sin perder prácticamente información de las p variables originales. Haciendo que dónde antes se necesitaban p valores para representar a un individuo ahora solo se necesiten z valores. Estos z valores se denominan componentes principales [7].

El análisis de componentes principales pertenece a la familia de técnicas conocida como *unsupervised learning*. El aprendizaje no supervisado es una clase de técnicas de *Machine Learning* para encontrar patrones en los datos. Los datos dados al algoritmo no supervisado no están etiquetados, lo que significa que sólo se dan las variables de entrada (X) sin las variables de salida correspondientes. En el aprendizaje no supervisado, los algoritmos se dejan a sí mismos para descubrir estructuras interesantes en los datos.

Los *eigenvectors*, también conocidos como vectores propios, son un caso particular de multiplicación de entre una matriz y un vector. El vector resultante de la multiplicación es un múltiplo entero del vector original. Los vectores propios de una matriz son todos aquellos vectores que, al multiplicarlos por dicha matriz, resultan en el mismo vector o en un múltiplo entero del mismo [8]. Los *eigenvectors* tienen un conjunto de propiedades matemáticas específicas:

- Los vectores propios solo existen para matrices cuadradas y no para todas. Sí una matriz $n \times n$ tiene *eigenvectors*, el número de estos será n.
- Si se escala un *eigenvector* antes de multiplicarlo por la matriz, se obtiene un múltiplo de este *eigenvector*. Esto se debe a que, si se escala un vector multiplicándolo por cierta cantidad, lo único que se consigue es cambiar su longitud no su dirección.

[7] De las siglas en inglés, Principal Component Analysis.

- Todos los vectores propios de una matriz son perpendiculares entre ellos, independientemente de las dimensiones que tengan.

Cuando se multiplica una matriz por alguno de sus vectores propios se obtiene un múltiplo del vector original, es decir, el resultado es ese mismo vector multiplicado por un número. Al valor por el que se multiplica el *eigenvector* se le denomina *egienvalue* o valor propio. A todo vector propio le corresponde un valor propio y viceversa.

En la técnica PCA, cada una de las componentes se corresponde con un *eigenvector*, y el orden de componente se establece por orden decreciente de *eigenvalue*. Así pues, la primera componente es el *eigenvector* con el *eigenvalue* más alto [9].

2.2.1.2. Cálculo de las componentes principales

Cada componente principal (Z_i) se obtiene por combinación lineal de las variables originales. Se pueden entender como nuevas variables obtenidas al combinar de una determinada forma las variables originales. La primera componente principal de un grupo de variables $(X_1, X_2, \dots, X_p)$ es la combinación lineal normalizada de dichas variables que tienen mayor varianza:

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p$$

Que la combinación lineal sea normalizada implica que:

$$\sum_{j=1}^{p} \phi_{j1}^2 = 1$$

Los términos $\phi_{11}, \dots, \phi_{p1}$ reciben el nombre de *loadings* y son los que definen a la componente. ϕ_{11} es el *loading* de la variable X_1 de la primera componente principal. Los *loaadings* pueden interpretarse como el peso/importancia que tiene cada variable en cada componente y, por lo tanto, ayudan a recoger que tipo de información recoge cada una de las variables [10].

Dado un set de datos X con n observaciones y p variables, el proceso a seguir para calcular la primera componente principal es:

- Normalización de las variables: se resta a cada valor el mínimo de la variable a la que pertenece y se divide por el rango de la variable, este, se calcula restándole el mínimo al máximo de la variable. Con esto se consigue que todas las variables un valor entre 0 y 1.
- Se resuelve un problema de optimización para encontrar el valor de los *loadings* con los que se maximiza la varianza. Una forma de resolver esta optimización es mediante el cálculo de *eigenvector-eigenvalue* de la matriz de covarianzas.

Una vez calculada la primera componente (Z_1) se calcula la segunda (Z_2) repitiendo el mismo proceso, pero añadiendo la condición de que la combinación lineal no puede estar correlacionada con la primera componente. Esto equivale a decir que las componentes Z_1 y Z_2 tienen que ser perpendiculares. El proceso se repite de forma iterativa hasta calcular todas las posibles componentes $(\min(n-1,\ p))$ o hasta que se decida detener el proceso. El orden de importancia de las componentes viene dado por la magnitud del valor propio asociado a cada vector propio [7].

Debido a que el método PCA identifica aquellas variables cuya variable es mayor. Es altamente recomendable estandarizar previamente las variables que se vayan a analizar, de tal manera que tengan media 0 i desviación estándar 1. Ya que, si no, a causa del cálculo de la variable que se mide en su misma escala elevada al cuadrado, aquellas variables cuya escala sea mayor dominarán el resto.

Para saber cuánta información se está perdiendo al representar todo el set de datos con una dimensión menor o, lo que es lo mismo, cuanta información es capaz de capturar cada una de las componentes principales, se utiliza a la proporción de varianza explicada por cada componente principal.

Presuponiendo la previa normalización de las variables para tener media cero, la presente varianza total en el set de datos se define de la siguiente manera:

$$\sum_{j=1}^{p} Var(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2$$

La varianza descrita por la componente *m* es:

$$\frac{1}{n} \sum_{i=1}^{n} z_{im}^2 = \frac{1}{n} \sum_{i=1}^{n} (\sum_{j=1}^{p} \phi_{jm} x_{ij})^2$$

Con lo cual, la proporción de varianza explicada por la componente *m* viene dada por la siguiente ratio:

$$\frac{\sum_{i=1}^{n}(\sum_{j=1}^{p}\phi_{jm}x_{ij})^2}{\sum_{j=1}^{p}\sum_{i=1}^{n}x_{ij}^2}$$

Estos dos últimos valores, varianza explicada y varianza explicada acumulada, son de gran utilidad para saber cuánta información se va a perder en la transformación. Si se utilizan todas las componentes de un set de datos, para hacer la conversión, se va a almacenar toda la información. Puesto que el sumatorio de la proporción de varianza explicada acumulada de todas las componentes es siempre 1.

Un ejemplo de la visualización de los datos después de la fusión de las X señales de características a 2 dimensiones sería la siguiente:

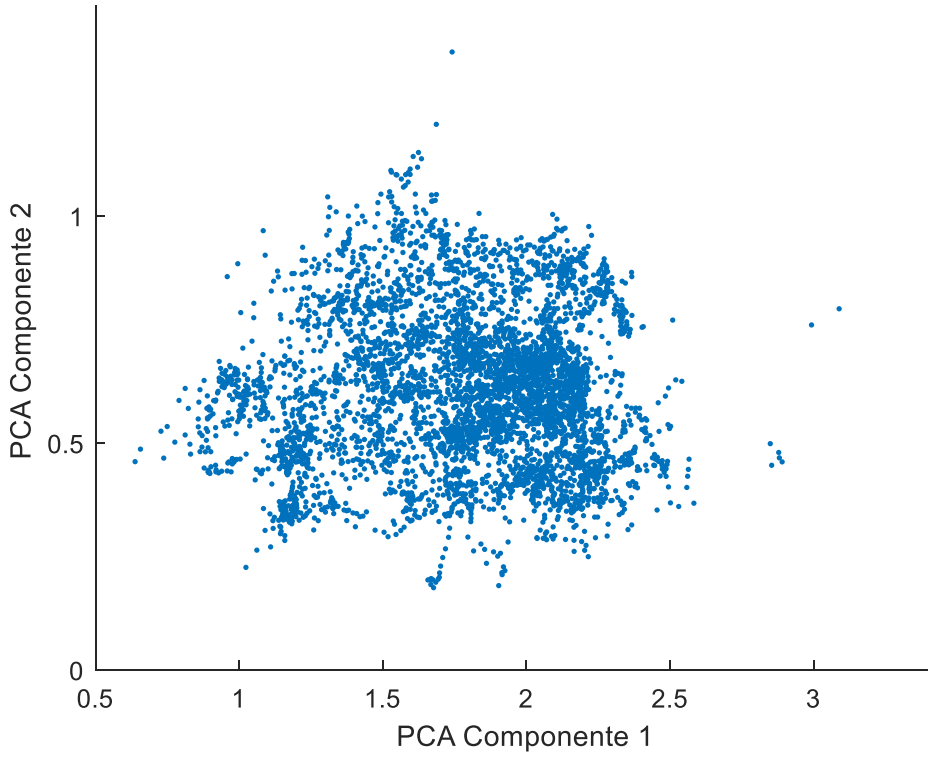

Ilustración 4. Ejemplo de visualización del resultado del método PCA

2.3. Modelado

A la salida del bloque anterior obtendremos un conjunto de puntos, que caracterizan cada una de las muestras de la sección, en un espacio de dos dimensiones. En este plano se podrá observar ciertos patrones en los datos, con una gran concentración de puntos en el plano. Esta concentración de puntos nos indicará el modo de operación normal. Llegado este punto, se requiere un algoritmo que permita delimitar estos espacios de operación para así poder detectar desviaciones. El método *One Class – Support Vector Machine* consigue resolver está necesidad. Esta técnica permite entrenar el modelo con datos de un funcionamiento correcto, en este caso de cada una de las secciones, con la intención de determinar cuándo una muestra se encuentra

fuera de esta normalidad. Se generará un segundo límite, menos restrictivo, para detectar los valores que se están desviando del proceso, sin llegar a ser una alarma.

2.3.1. One Class - Support Vector Machine (vSVM)

Este método está pensado para dividir el conjunto de datos en varias clases, en este caso solo tenemos una clase, que será el estado de un correcto funcionamiento de la sección. Pero para poder entender cómo funciona el modelo se explicará cómo se realiza para dos clases distintas.

Considerando un conjunto de datos $\Omega = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$; puntos $x_i \epsilon \mathbb{R}^d$ en un espacio donde x_i es el punto de entrada de datos i-ésimo i $y_i \epsilon \{-1,1\}$ es el i-ésimo patrón de salida, indicando así la clase a la que pertenece. Una de las propiedades del SVM es que puede crear un límite no-lineal proyectando los datos a través de una función no-lineal ϕ a un espacio de una dimensión mayor. Esto significa que los puntos de datos que no pueden ser divididos por una línea recta en su espacio original I son "elevados" a un espacio de características F donde puede haber un hiperplano "recto" que separa los puntos de datos de una clase de otra. Una vez proyectado de vuelta al espacio de entrada I, estos tendrán una forma de curva no-linear [11].

El hiperplano se representa con la ecuación $w^t x + b = 0$, con $w \epsilon F$ i $b \epsilon R$. El hiperplano construido es quien determina la frontera entre las clases. Todos los puntos de datos de la clase -1 quedan a un lado del hiperplano, y todos los puntos de datos de la clase 1 en el otro. La distancia entre el punto más cercano de cada clase al hiperplano es la misma. Con lo cual, el hiperplano construido busca el límite máximo entre las clases. Para prevenir que el clasificador SVM sufra *over-fitting*[8] con datos con ruido (ruido blanco estadístico), o para crear márgenes más suaves, se añaden al modelo las variables de frontera ξ_i para permitir que algunas muestras de una clase se encuentren dentro de la frontera, y la constante $C > 0$ determina el compromiso entre maximizar el límite y el número de puntos de datos de formación dentro de esa frontera. La función que caracteriza el clasificador SVM es la siguiente:

[8] Sobreajuste, es el efecto de sobreentrenar un algoritmo de aprendizaje con unos ciertos datos para los que se conoce el resultado deseado

$$\min_{w,b,\xi_i} \frac{\|w\|^2}{2} + C\sum_{i=1}^{n} \xi_i$$

Tal que:

$$y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i \quad para\ todo\ i = 1, \dots, n$$

$$\xi_i \geq 0 \quad para\ todo\ i = 1, \dots, n$$

Al resolver este problema de minimización, utilizando multiplicadores de Lagrange, la función de decisión o clasificación de la regla para un punto de datos x se convierte entonces en la siguiente fórmula:

$$f(x) = sgn\left(\sum_{i=1}^{n} a_i y_i K(x, x_i) + b\right)$$

En esta expresión α_i son los multiplicadores de Lagrange. Cada $\alpha_i > 0$ es ponderado en la función de decisión. Dado que se considera que los SVMs son escasos, habrá relativamente pocos multiplicadores de Lagrange con un valor distinto de cero.

La función $K(x, x_i) = \phi(x)^T\phi(x_i)$, también conocida como *Kernel Function*. Puesto que el resultado de la función de decisión sólo se basa en el producto de puntos de los vectores en el espacio de características F, no es necesario realizar una proyección explícita a ese espacio. Mientras la función K tenga los mismos resultados, se puede utilizar en su lugar. Esto se conoce como el *kernel trick* o truco de núcleo y es lo que da a la técnica SVM una gran potencia con puntos de datos no separables linealmente. El espacio de características F puede ser de dimensión ilimitada, con lo cual, el hiperplano que separa los datos puede ser muy complejo. Sin embargo, en nuestros cálculos evitamos esa complejidad.

Las opciones más populares para la función de kernel son lineal, polinómica, sigmoidal, pero sobre todo la función de base radial gaussiana (RBF)[9]:

$$K(x, x') = exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$$

dónde $\sigma \epsilon R$ es un parámetro del núcleo y $\|x - x'\|$ es la mesura de diferencia.

[9] De las siglas en inglés *Gaussian Radial Base Function*

Con este conjunto de fórmulas y conceptos podemos clasificar un conjunto de puntos de datos en dos clases con una función de decisión no lineal. Para este estudio se requiere de las técnicas para resolver el problema con una sola clase, *One-Class Support Vector*. El procedimiento de esta técnica consiste en separar todos los puntos de datos del origen, en el espacio de características F, y maximizar la distancia entre este hiperplano hasta el origen. Esto tiene como resultado una función binaria que captura regiones en el espacio de entrada donde se concentra la densidad de probabilidad de los datos. De este modo, la función devuelve +1 en una región pequeña (capturando los datos de entrenamiento) y -1 en otra parte [12].

La función de minimización de la función de programación cuadrática es ligeramente diferente a la original mencionada anteriormente, pero la similitud sigue siendo clara:

$$\min_{w,\xi_i,\rho} \frac{1}{2}\|w\|^2 + \frac{1}{vn}\sum_{i=1}^{n}\xi_i - \rho$$

Tal que:

$$\big(w \cdot \phi(x_i)\big) \geq \rho - \xi_i \qquad para\ todo\ i = 1, \dots, n$$

$$\xi_i \geq 0 \qquad para\ todo\ i = 1, \dots, n$$

En la formula anterior, el parámetro C decidía la suavidad. En Esta fórmula es el parámetro v el que caracteriza la solución:

- Fija un límite superior en la fracción de *outliers*.
- Establece un límite inferior para el número de muestras de entrenamiento utilizados como vectores de soporte.

A causa de este parámetro a esta técnica se le puede llamar, *v-SVM*.

Una vez más, utilizando las técnicas de Lagrange y una función de *kernel* para el cálculo de productos puntuales, se convierte en la función de decisión:

$$f(x) = sgn\Big(\big(w \cdot \phi(x_i)\big) - \rho\Big) = sgn\left(\sum_{i=1}^{n}\alpha_i \mathrm{K}(x, x_i) - \rho\right)$$

Este método crea así un hiperplano caracterizado por w y ρ que tiene una distancia máxima desde el origen al espacio de características F y separa todos los puntos de datos del origen.

En la siguiente ilustración se muestra un ejemplo de los límites calculados con el método *One-Class Support Vector Machine*.

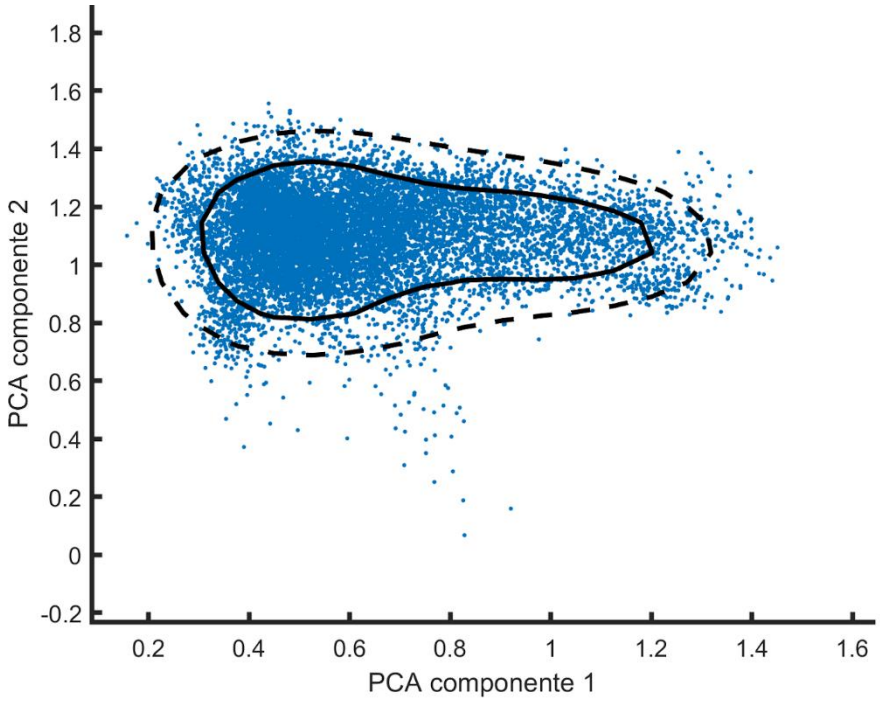

Ilustración 5. Ejemplo de visualización de los límites calculados con vSVM

2.3.2. Índice de normalidad

Una vez delimitada la frontera de normalidad y desviación utilizando el método *vSVM* se debe evaluar los datos frente a estos límites. Para ello vamos a asignar 3 valores distintos en función del espacio en el que se encuentre cada una de las muestras, llamado índice de normalidad. Este indicé tiene los siguientes valores:

- 2: Si se encuentra dentro de los límites de normalidad
- 1: Si se encuentra entre el límite de normalidad y el límite de desviación
- 0: Si se encuentra fuera del límite de desviación

Para cada sección tendremos un vector de normalidad, del tamaño igual al número de muestras, dónde se incluirá el índice de normalidad de cada muestra. Este índice nos indicara cuando una muestra se está alejando de la zona de normalidad del proceso.

2.3.3. Índice de normalidad global

Por último, se calculará el índice de normalidad global, la suma de cada índice de normalidad por sección. Este índice nos dará información de cuando el proceso se está desviando de la normalidad conjunta del proceso de laminación. Aquellas muestras que tiendan a comportarse fuera de la normalidad, tanto global como individual, por sección, van a tener más posibilidades de que ocurra un fallo. Esto se comprobará analizando un conjunto de 95 palanquillas, que son fallos, con esta metodología. Comparándola, finalmente, con el primer conjunto de datos.

2.4. Visualización online

Para que todo lo visto anteriormente tenga una utilidad práctica se debe poder visualizar a tiempo real los datos, para así poder detectar desviaciones del proceso y tomar las decisiones oportunas para corregirlo. Para la visualización de los datos se va a utilizar una herramienta de *Business Intelligence* llamada Tableau. Para ello, será necesario que los datos, extraídos del modelo, se introduzcan a una BBDD a tiempo real para posteriormente extraer los datos para realizar la visualización.

La visualización que más va a interesar para la detección de fallos será la evolución en el tiempo del índice de normalidad global, así como, la afectación de cada una de las secciones a este índice. De esta manera, los técnicos de la planta al detectar un índice de normalidad global bajo van a poder detectar cuál de las secciones es la afectada y hacer un análisis para la posterior reconfiguración de los parámetros o mantenimiento de la sección.

3. RESULTADOS EXPERIMENTALES

3.1. Preprocesado

3.1.1. Exploración de datos en crudo

Tal y como se mencionó previamente el objetivo de esta exploración es la de identificar rangos de operación normal de las señales para posteriormente poder limpiar las señales de *outliers* y mal registros. En el siguiente apartado se puede ver los resultados de la primera sección, horno de recalentamiento, los resultados de las secciones siguientes se pueden encontrar en el anexo.

3.1.1.1. Horno de recalentamiento

La matriz de indicadores de la primera sección del proceso está formada por 15 sensores distintos. En las siguientes imágenes se puede ver la representación de los números de muestras de las señales por cada uno de los sensores pertenecientes a esta sección.

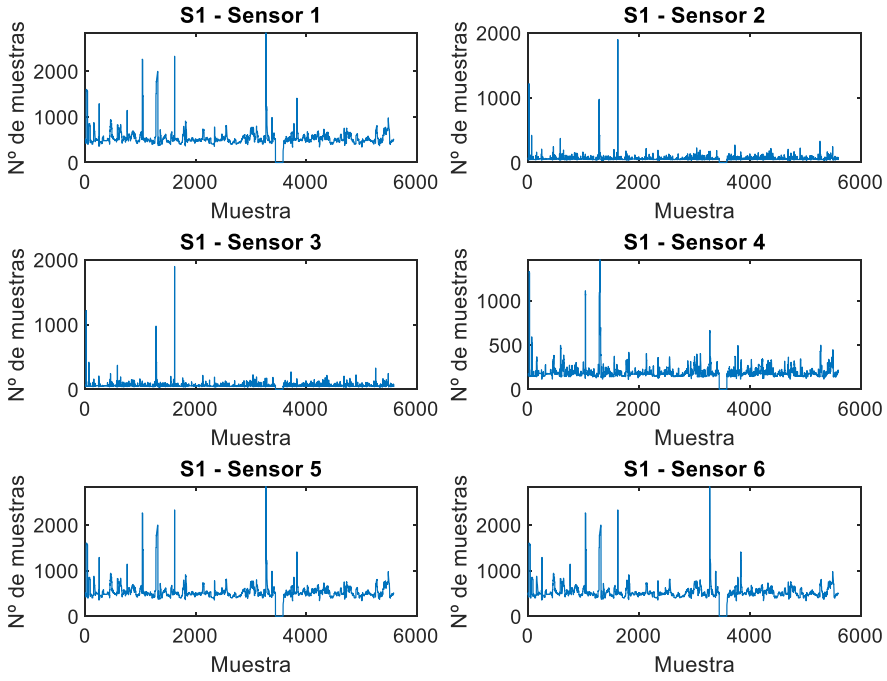

Ilustración 6. Número de muestras de la señal por sensores de la Sección 1 I

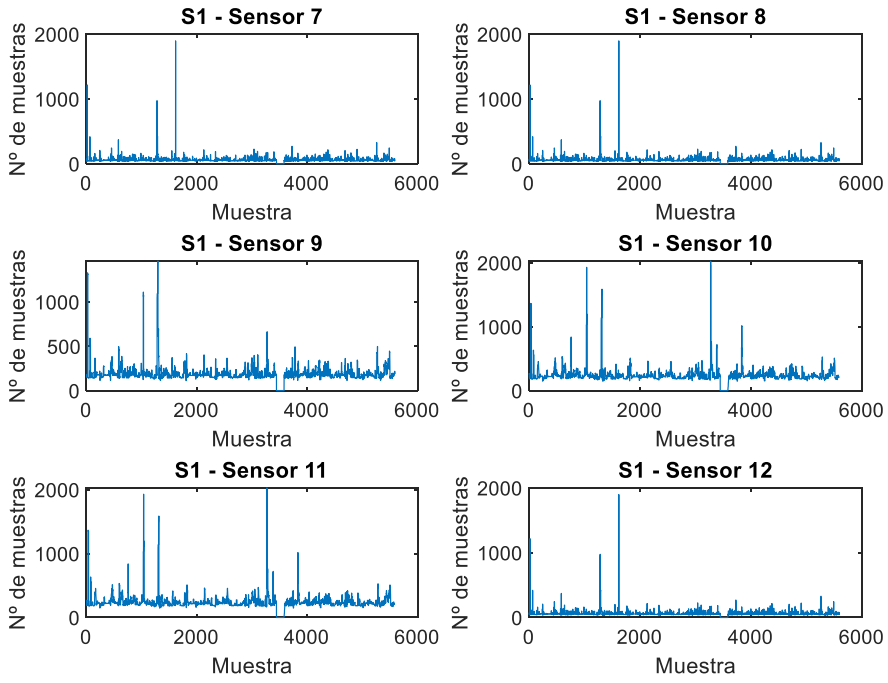

Ilustración 7. Número de muestras de la señal por sensores de la sección 1 II

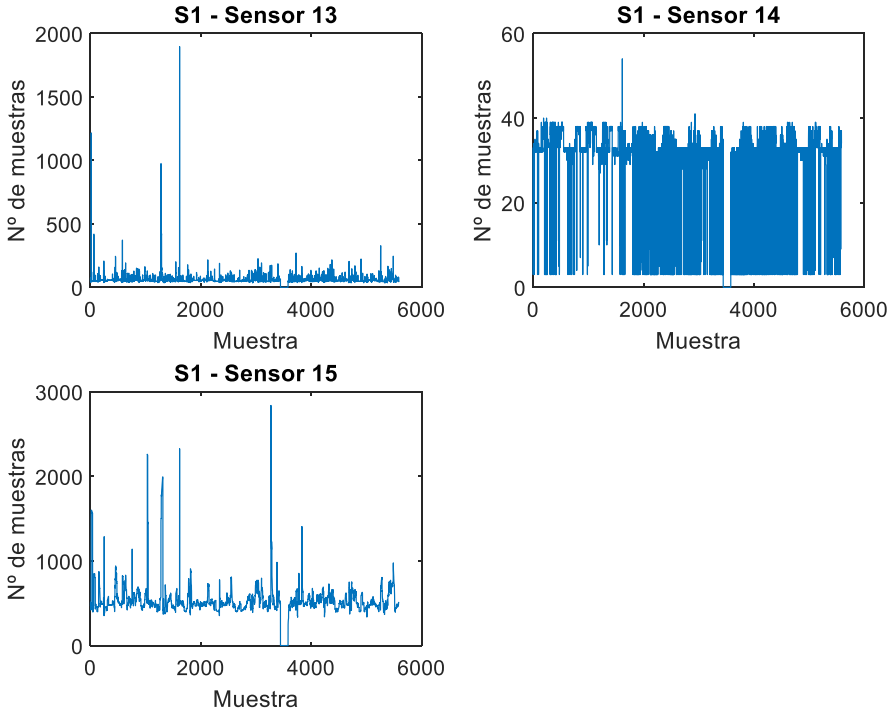

Ilustración 8. Número de muestras de la señal por sensores de la sección 1 III

Como se puede observar, excepto el sensor 14, todos los sensores tienen registros con números de muestras muy distintos a la media de número de muestras por los sensores. Alrededor las muestras 3400 hasta 3600 podemos observar que hay una zona de valores igual a 0. En las siguientes imágenes se muestran los valores de la media de las señales por cada sensor:

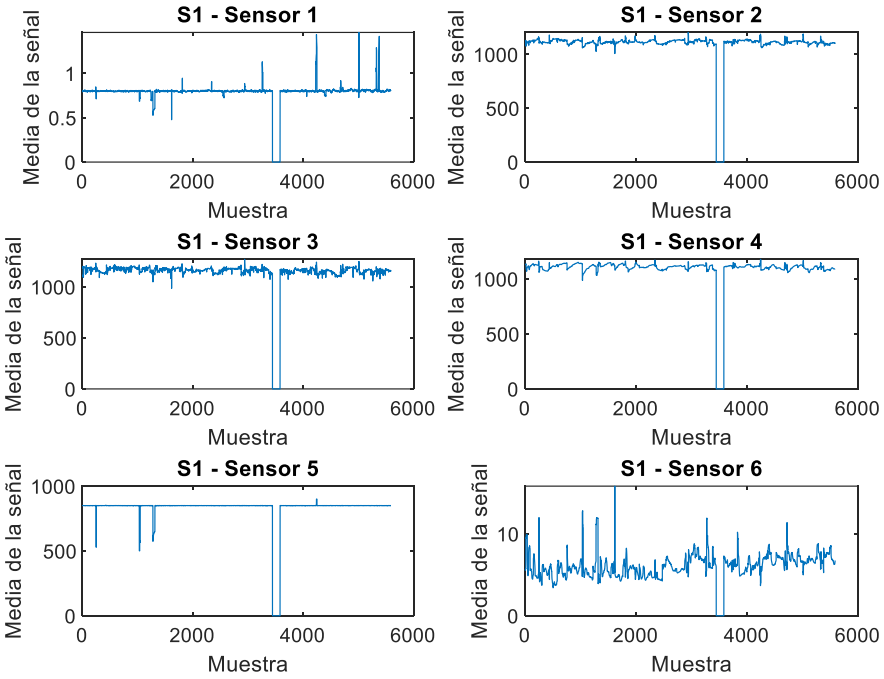

Ilustración 9. Valor medio de las señales por sensores de la Sección 1 I

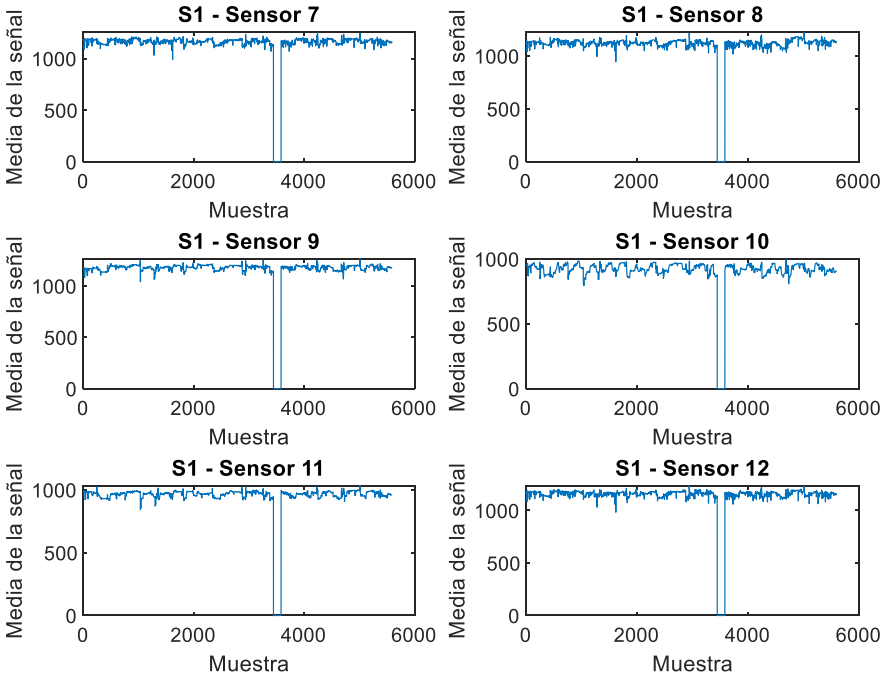

Ilustración 10. Valor medio de las señales por sensores de la sección 1 II

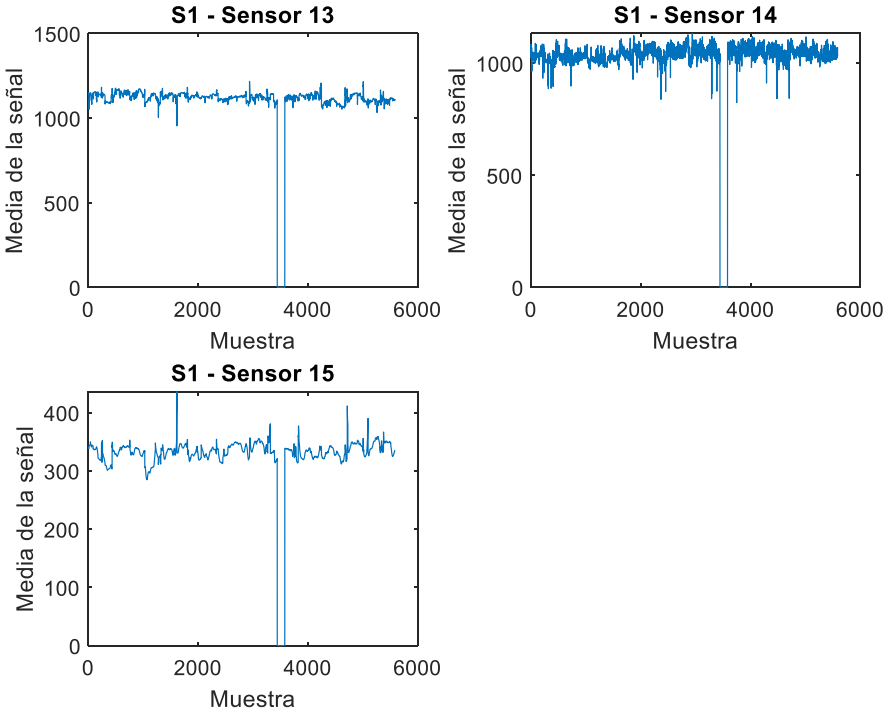

Ilustración 11. Valor medio de las señales por sensores de la sección 1 III

En el caso de las medias de la señal, también se puede contemplar que en el rango de muestras de 3400 a 3600 el valor de media de la señal es igual a 0. Por lo demás las medias de las señales parecen tener un rango muy estable.

Como conclusión, se ha podido observar que a partir de la sección 2 aparecen una cantidad considerables de registros defectuosos, que se van a tener que filtrar para el entrenamiento del modelo. Si bien es cierto que en la sección 1 también aparece algún mal registro, pero no tan destacable como en las otras secciones.

3.1.2. Limpieza de datos

Una vez identificado los rangos de operación normal de cada una de las señales del proceso, se procede a eliminar los registros que no cumplan con los criterios previamente establecidos. Igual que en la sección anterior, solo se va a mostrar los resultados de la primera sección, los resultados de las siguientes secciones se pueden visualizar en el anexo.

3.1.2.1. Horno de recalentamiento

A esta sección el primer filtro que se le ha aplicado es sobre el número de muestras del primer sensor con un límite inferior a 300 números de muestra y un límite superior de 900, descartando un total de 604 palanquillas (10% de las barras totales). El segundo filtro que se ha empleado son todas aquellas muestras que tengan la mitad de las medias de las señales de los diferentes sensores de una sección igual a cero, en este caso no ha habido ninguna muestra que lo cumpliera. Por último, se ha utilizado un filtro para eliminar los *outliers*, en este caso todos aquellos registros de la media de la señal del sensor uno superior a 0.84 e inferior a 0.77 han sido descartados, eliminando un total de 337 muestras (5% de las barras totales).

En los siguientes gráficos se muestran las señales después de la aplicación de los filtros:

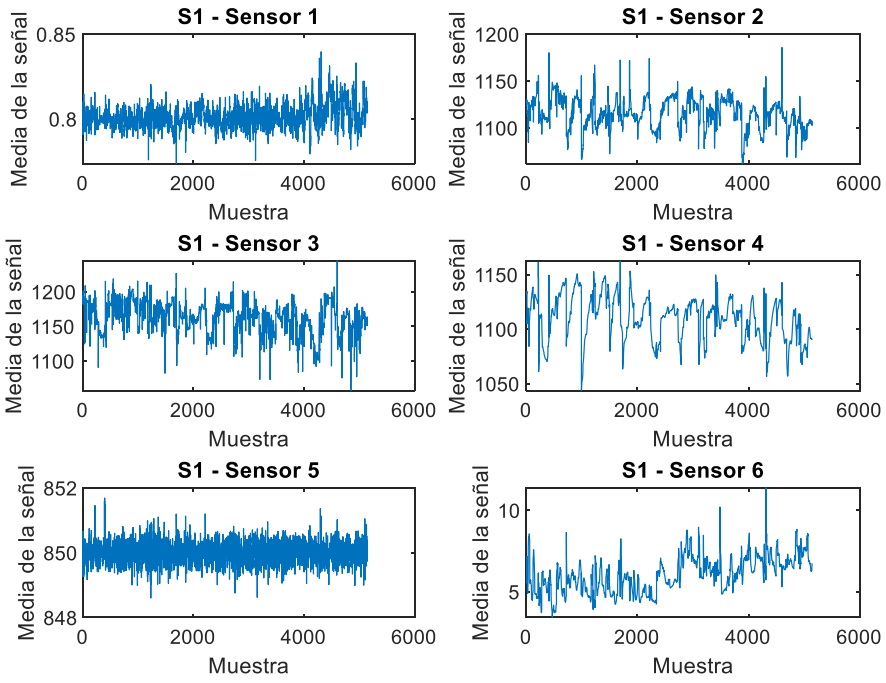

Ilustración 12. Valores filtrados de la media de las señales por sensores de la sección 1 I

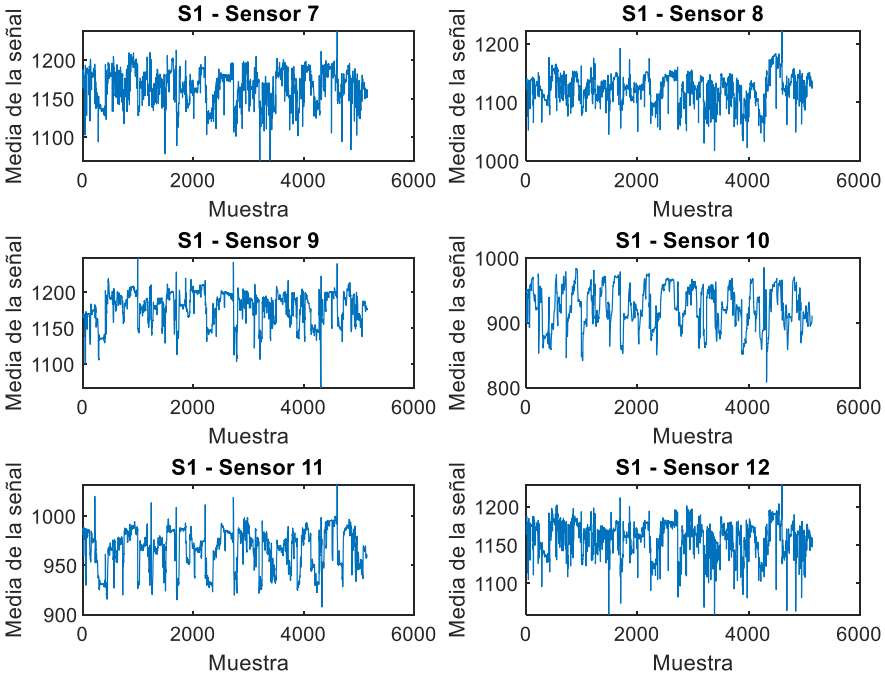

Ilustración 13. Valores filtrados de la media de las señales por sensores de la sección 1 II

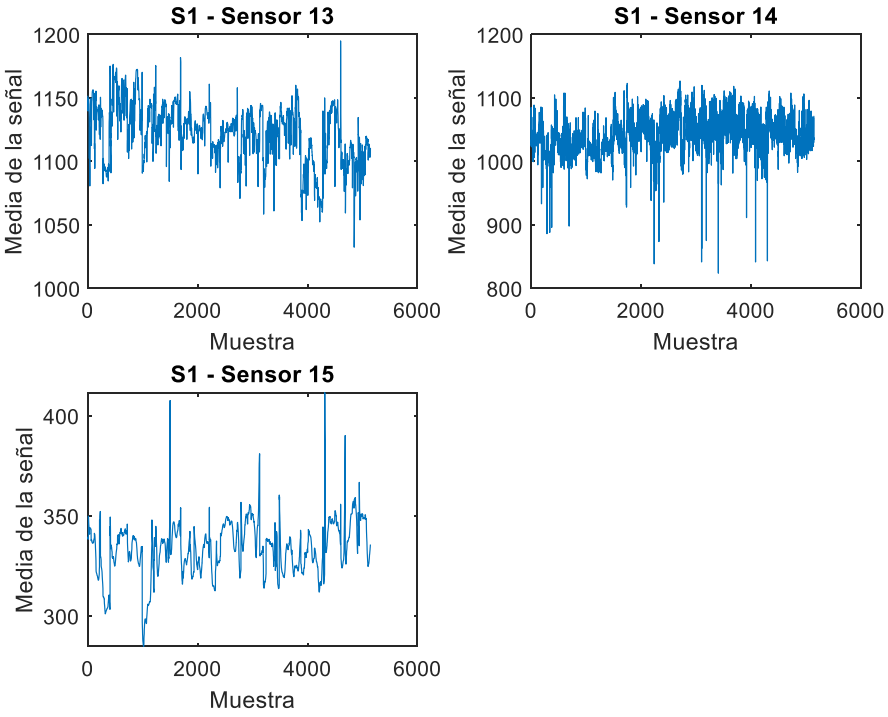

Ilustración 14. Valores filtrados de la media de las señales por sensores de la sección 1 III

Se ha podido ver en las gráficas anteriores que gracias a los filtros aplicados a las señales se ha conseguido eliminar aquellas muestras mal registradas. Además, se ha conseguido eliminar, también, los valores anómalos de las señales que podrían llegar a influir negativamente al resultado una vez aplicado el modelo. Después de la aplicación de los filtros se ha conseguido una limitación de la variabilidad de las señales.

De la misma manera, se ha podido observar que el segundo filtro que se ha aplicado en todas las secciones no se ha cumplido en ningún caso, eso es debido a dos casuísticas, la sección no tenía ningún valor que cumpliera la regla o el primer filtro utilizado ha conseguido descartar todos aquellos valores que cumplieran la condición.

3.2. Procesado para caracterizar patrones

Como ya se ha comentado en el apartado 2.1.2. la manera de calcular los componentes principales es haciendo una normalización de las variables, haciendo que el rango de las variables sea 1 y el valor mínimo sea 0. Una vez se tienen las variables normalizadas, se procede al cálculo de la matriz de covarianzas de las variables, que servirán para el posterior cálculo de las componentes principales de cada variable. Por último, se hará el cálculo de la ratio de la varianza explicada, que servirá para explicar cuanta información se está perdiendo. Los resultados de todas las secciones se encuentran en el anexo.

3.2.1. PCA

3.2.1.1. Horno de recalentamiento

La matriz de covarianzas de esta sección una vez se han normalizado las señales, es la siguiente:

Comp.	C 12	C2	C3	C4	C5	C6	C7
C1	0,009279	-0,000391	0,000055	-0,000929	-0,002371	0,002268	0,000969
C2	-0,000391	0,016525	0,010638	0,014992	0,000513	-0,004358	0,011538
C3	0,000055	0,010638	0,014468	0,010574	0,000692	-0,003583	0,012852
C4	-0,000929	0,014992	0,010574	0,028330	-0,000284	-0,006166	0,010249
C5	-0,002371	0,000513	0,000692	-0,000284	0,009601	-0,000401	0,000507
C6	0,002268	-0,004358	-0,003583	-0,006166	-0,000401	0,017966	-0,000942
C7	0,000969	0,011538	0,012852	0,010249	0,000507	-0,000942	0,016323
C8	0,001755	0,010724	0,011985	0,008377	0,000577	-0,002203	0,012744
C9	0,000369	0,008493	0,008870	0,010663	-0,000303	-0,003300	0,010065
C10	-0,000245	0,013223	0,010859	0,022794	-0,000763	-0,003456	0,011442
C11	0,000448	0,009170	0,010808	0,013721	0,000168	-0,000788	0,012251
C12	0,000884	0,011351	0,013899	0,009883	0,000697	-0,001657	0,014838
C13	-0,000717	0,013223	0,014103	0,015448	0,000395	-0,008373	0,011382
C14	-0,000117	0,003157	0,002811	0,000584	0,000228	0,001732	0,004336
C15	0,000637	0,003687	0,003517	0,005199	-0,000537	0,005932	0,004556

Tabla 1. Matriz de covarianza de la S1 I

Comp.	C8	C9	C10	C11	C12	C13	C14	C15
C1	0,00175	0,00037	-0,00025	0,00045	0,00088	-0,00072	-0,00012	0,00064
C2	0,01072	0,00849	0,01322	0,00917	0,01135	0,01322	0,00316	0,00369
C3	0,01199	0,00887	0,01086	0,01081	0,01390	0,01410	0,00281	0,00352
C4	0,00838	0,01066	0,02279	0,01372	0,00988	0,01545	0,00058	0,00520
C5	0,00058	-0,00030	-0,00076	0,00017	0,00070	0,00039	0,00023	-0,00054
C6	-0,00220	-0,00330	-0,00346	-0,00079	-0,00166	-0,00837	0,00173	0,00593
C7	0,01274	0,01006	0,01144	0,01225	0,01484	0,01138	0,00434	0,00456
C8	0,01440	0,00910	0,01094	0,01037	0,01358	0,01173	0,00284	0,00478
C9	0,00910	0,01495	0,01059	0,01276	0,01009	0,00969	0,00441	0,00481
C10	0,01094	0,01059	0,03151	0,01939	0,01131	0,01242	0,00083	0,00852
C11	0,01037	0,01276	0,01939	0,02515	0,01211	0,00979	0,00472	0,00708
C12	0,01358	0,01009	0,01131	0,01211	0,01576	0,01232	0,00416	0,00464
C13	0,01173	0,00969	0,01242	0,00979	0,01232	0,01926	0,00191	0,00235
C14	0,00284	0,00441	0,00083	0,00472	0,00416	0,00191	0,00869	0,00208
C15	0,00478	0,00481	0,00852	0,00708	0,00464	0,00235	0,00208	0,01102

Tabla 2. Matriz de covarianza de la S1 II

Está matriz será la que utilizaremos como entrada del método PCA, esté modelo nos dará las componentes principales y la varianza explicada. En la siguiente tabla podemos ver cuáles son las PC:

Comp.	C1	C2	C3	C4	C5	C6	C7
C1	0,00304	0,16379	-0,01728	-0,20812	-0,61887	-0,12583	0,65745
C2	0,29057	-0,11503	-0,10784	-0,28960	0,13305	0,22219	0,10646
C3	0,28414	0,03162	-0,27786	-0,06901	0,02198	-0,16264	-0,14712
C4	0,36278	-0,35172	0,37095	-0,27527	0,08791	0,33584	0,28382
C5	0,00438	-0,02405	-0,08885	0,14071	0,70420	-0,29420	0,55370
C6	-0,08793	0,65189	0,21026	-0,40153	0,20643	0,08264	0,05883
C7	0,29444	0,19822	-0,24801	-0,06465	0,01589	-0,07026	0,02866
C8	0,27131	0,14206	-0,26166	-0,10930	-0,07355	-0,24938	-0,05133
C9	0,25431	0,09284	-0,06450	0,32445	-0,12018	0,41574	0,07287
C10	0,38955	-0,07650	0,56211	-0,01455	-0,02971	-0,35920	-0,14037

C11	0,33409	0,26109	0,24218	0,62196	-0,08019	-0,11145	0,09456
C12	0,29775	0,17703	-0,28566	-0,05617	0,00974	-0,13818	-0,05118
C13	0,31232	-0,25939	-0,25250	-0,13283	-0,00237	0,05599	-0,11645
C14	0,07216	0,22976	-0,12721	0,20935	0,11031	0,53562	0,08849
C15	0,12540	0,34085	0,22553	-0,19543	0,11352	0,10792	-0,28769

Tabla 3. Matriz de componentes principales de la S1 I

Comp.	C8	C9	C10	C11	C12	C13	C14	C15
C1	0,10567	-0,10826	0,21629	-0,05290	0,06047	-0,14050	-0,04988	0,01088
C2	-0,31358	-0,29390	0,02302	0,66198	-0,15653	-0,19224	-0,20626	-0,03115
C3	0,04502	0,28108	0,13606	-0,22329	0,02009	-0,19660	-0,52922	-0,56497
C4	0,05034	0,31211	-0,16367	-0,13823	0,27931	0,31961	-0,01889	-0,06230
C5	0,22213	-0,13633	0,07405	-0,06826	-0,00967	-0,06657	0,00558	0,00756
C6	-0,05993	0,35553	-0,05578	-0,01449	-0,41621	0,02684	0,04022	0,00940
C7	-0,23323	0,03626	-0,41135	-0,09279	0,33222	-0,37318	0,54903	-0,13312
C8	0,09536	-0,26138	-0,01602	0,09008	-0,13011	0,73298	0,22016	-0,25908
C9	0,48355	-0,16873	-0,42251	-0,07898	-0,39179	-0,13040	-0,07967	-0,02802
C10	-0,20196	-0,35339	0,03233	-0,31506	-0,28525	-0,16710	0,00928	0,03546
C11	-0,09831	0,35485	0,18191	0,40393	0,09774	0,07120	0,02170	0,00314
C12	-0,08498	0,03790	-0,17097	-0,14946	0,15395	0,13435	-0,39415	0,71840
C13	0,23966	0,28769	0,50189	-0,03647	-0,29847	-0,15754	0,40000	0,26410
C14	-0,39912	-0,25523	0,41242	-0,39993	0,05901	0,11736	0,02650	-0,05177
C15	0,51116	-0,28680	0,25586	0,12704	0,48178	-0,11911	-0,00519	0,01626

Tabla 4. Matriz de componentes principales de la S1 II

En la siguiente gráfica podremos ver la cantidad de información acumulad0a de cada una de las componentes:

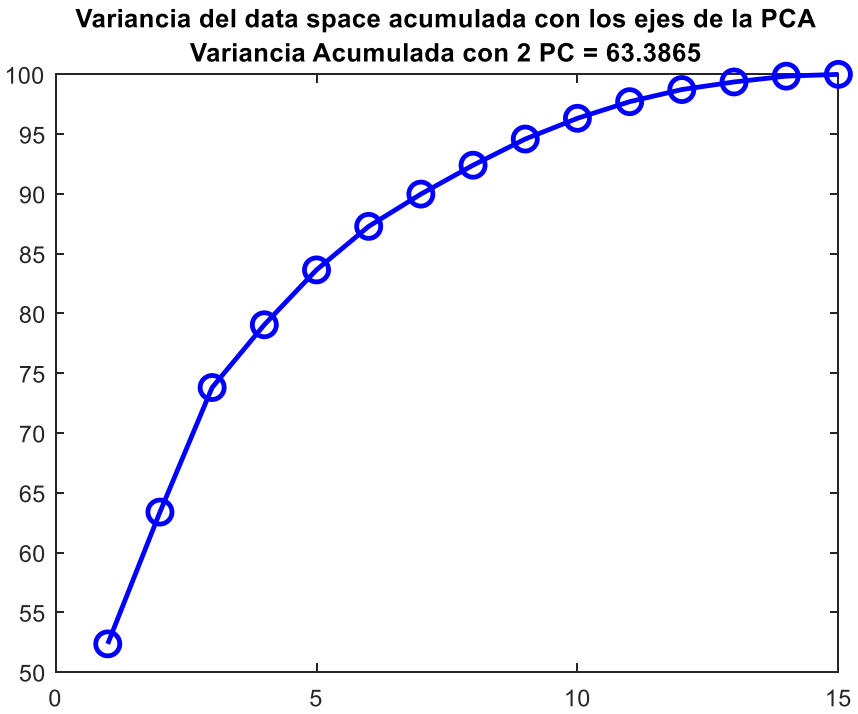

Ilustración 15. Variancia acumulada por componentes de la PCA de la S1

Lo que nos indica está gráfica es que con solo 2 componentes principales estamos obteniendo un 63.4% de la información que obtendríamos con las 15 componentes. Veamos pues como queda la matriz de componentes una con solo las dos componentes principales:

Componente 1	Componente 2
0,00304	0,16379
0,29057	-0,11503
0,28414	0,03162
0,36278	-0,35172
0,00438	-0,02405
-0,08793	0,65189
0,29444	0,19822
0,27131	0,14206
0,25431	0,09284
0,38955	-0,07650
0,33409	0,26109
0,29775	0,17703

0,31232	-0,25939
0,07216	0,22976
0,12540	0,34085

Tabla 5. Valores de las 2 componentes principales con más información de la S1

Una vez obtenida la matriz de covarianzas se procede a realizar una multiplicación de matrices, la matriz de indicadores filtrada con la PC reducida, esto nos dará un resultado de una matriz de dos dimensiones con 5144 registros, uno por cada palanquilla no filtrada. Estás dos dimensiones serán usadas para representarlas en un plano. En la siguiente imagen se puede observar está representación:

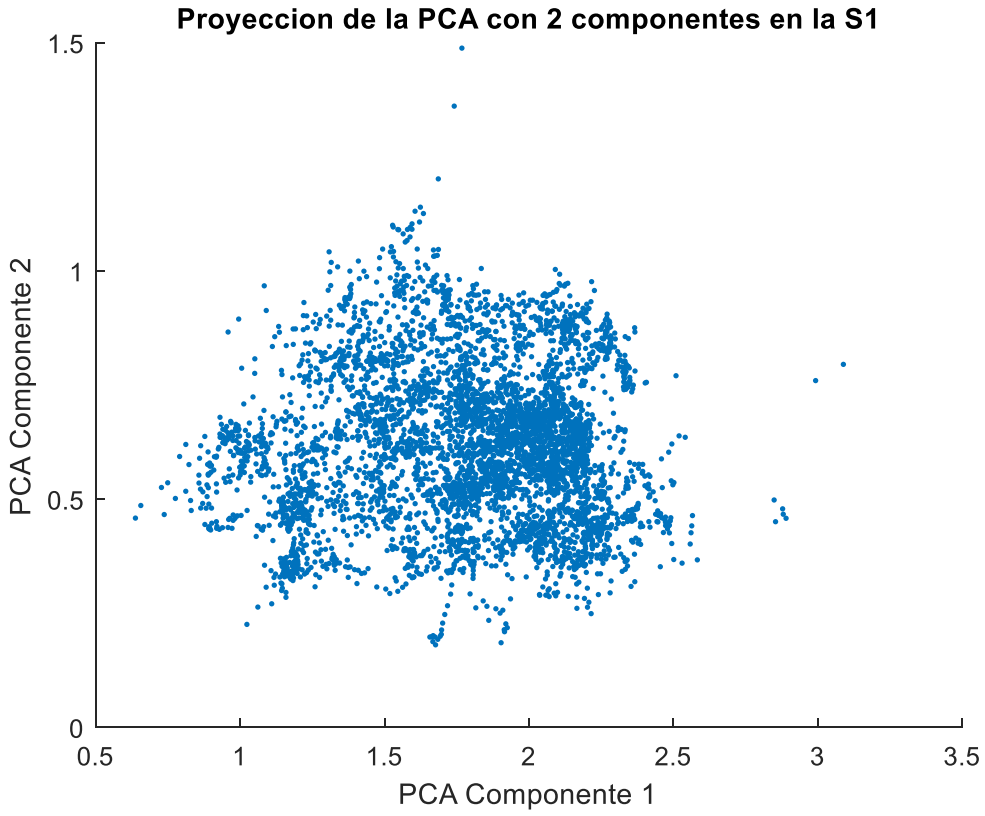

Ilustración 16. Proyección de las muestras filtradas de la S1 sobre las 2 componentes principales de la PCA

3.3. Modelado

3.3.1. One-Class Support Vector Machine

Una vez ya se tiene la reducción de dimensionalidad hecha y se puede ver como se distribuyen los puntos de datos en este nuevo plano, aparece la necesidad de determinar qué zonas de puntos están trabajando dentro de la normalidad del proceso. Como ya se ha explicado en el apartado 2.3.1. el método *One-Class Support Vector Machine* es una opción adecuada para realizar esta tarea.

En primer lugar, se van a cargar todos los datos sin filtrar y se hará el cambio de base utilizando los componentes principales calculados en el apartado anterior por cada una de las secciones. Seguidamente, se fijará un límite a un 10% de datos que quedarán fuera de la normalidad, este límite es fijado para dejar fuera a los *outliers*. La siguiente acción es configurar el valor *v*, encargado de ajustar el límite a los datos requeridos. Cabe remarcar que un valor muy bajo puede dar pie a un sobreajuste de los datos, llamado *overffiting*. Este parámetro se ajusta en base al criterio del usuario, teniendo en cuenta la visualización del límite, considerando una compensación entre la precisión del límite y un *overfitting*. Para finalizar se va a calcular el valor obtenido de datos que quedan fuera del límite de normalidad, para así comprobar lo ajustado que queda el límite con el parámetro utilizado.

3.3.1.1. Horno de recalentamiento

Para la sección uno se ha utilizado una configuración del parámetro *v* igual a 0.45, el resultado del límite de normalidad se puede observar en la siguiente ilustración.

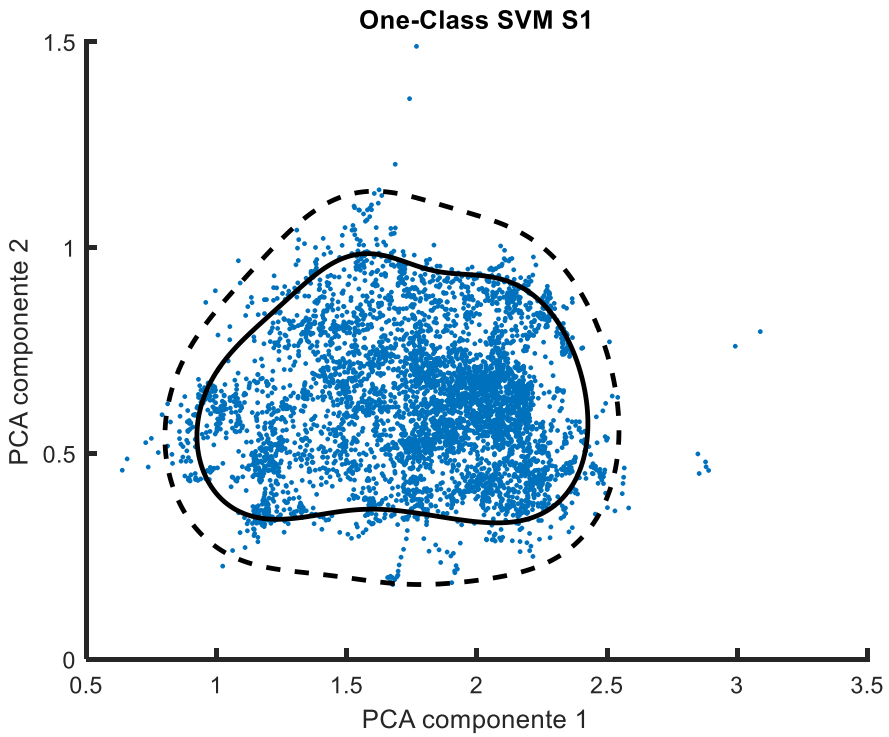

Ilustración 17. Límites de normalidad con One-Class SVM S1

Con la configuración establecida se ha conseguido descartar exactamente el 10% de los datos.

3.3.1.2. Tren de desbaste

La sección dos se ha configurado con un parámetro *v* de 0.85. En la siguiente imagen se muestra como han quedado los límites.

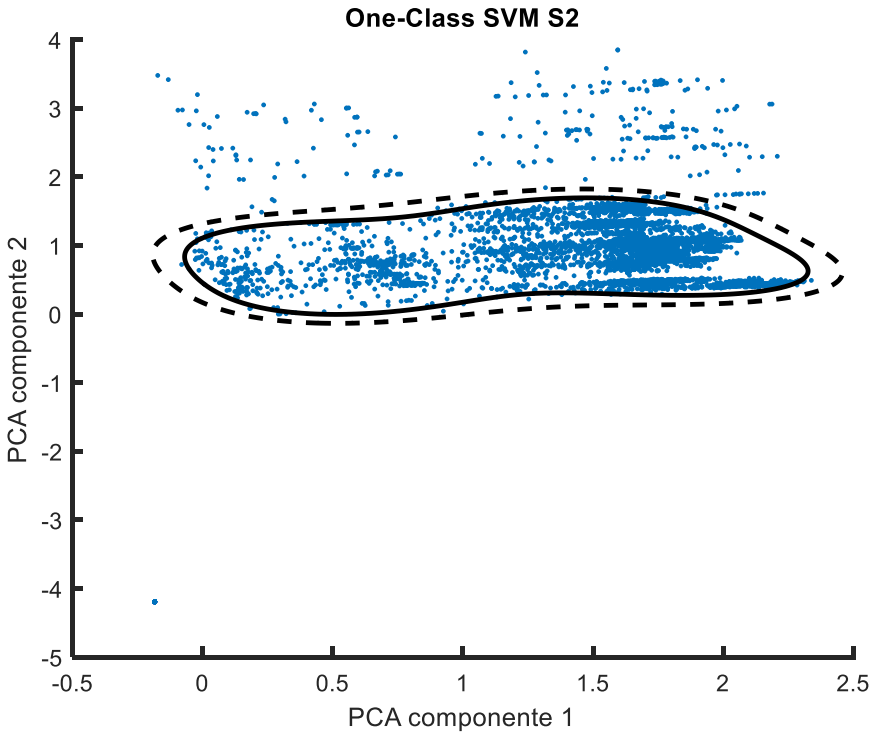

Ilustración 18. Límites de normalidad con One-Class SVM S2

Con la configuración establecida quedan fuera de los límites de normalidad un 14% de los datos.

3.3.1.3. Tren intermedio

El parámetro de configuración *v* de la sección 3, se ha ajustado a 0.9. El resultado de esta configuración se puede ver en el siguiente gráfico.

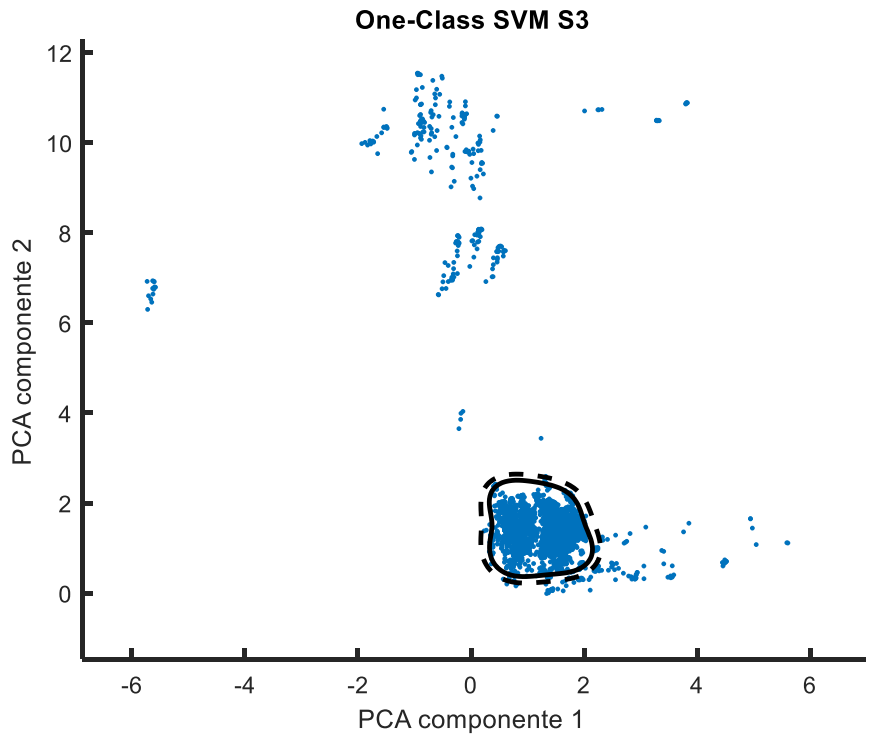

Ilustración 19. Límites de normalidad con One-Class SVM S3

Los resultados de esta configuración han dejado fuera de los límites un 14.5% del total de los datos.

3.3.1.4. Tren acabador

El parámetro de configuración *v* de la sección 4 es igual a 0.3, el resultado de tal configuración sobre los datos de esta sección es el siguiente.

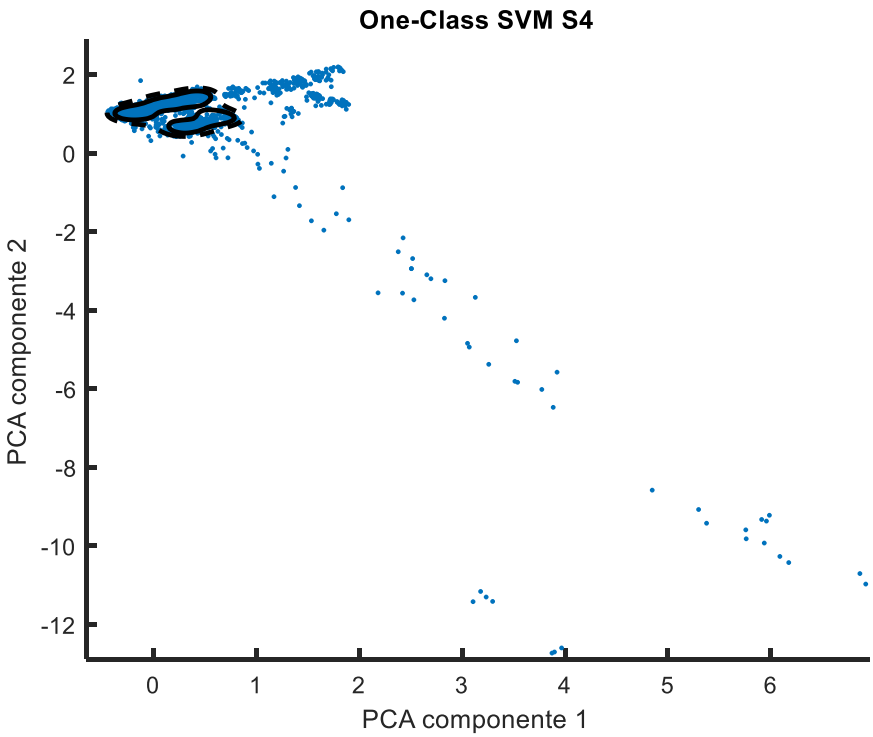

Ilustración 20. Límites de normalidad con One-Class SVM S4

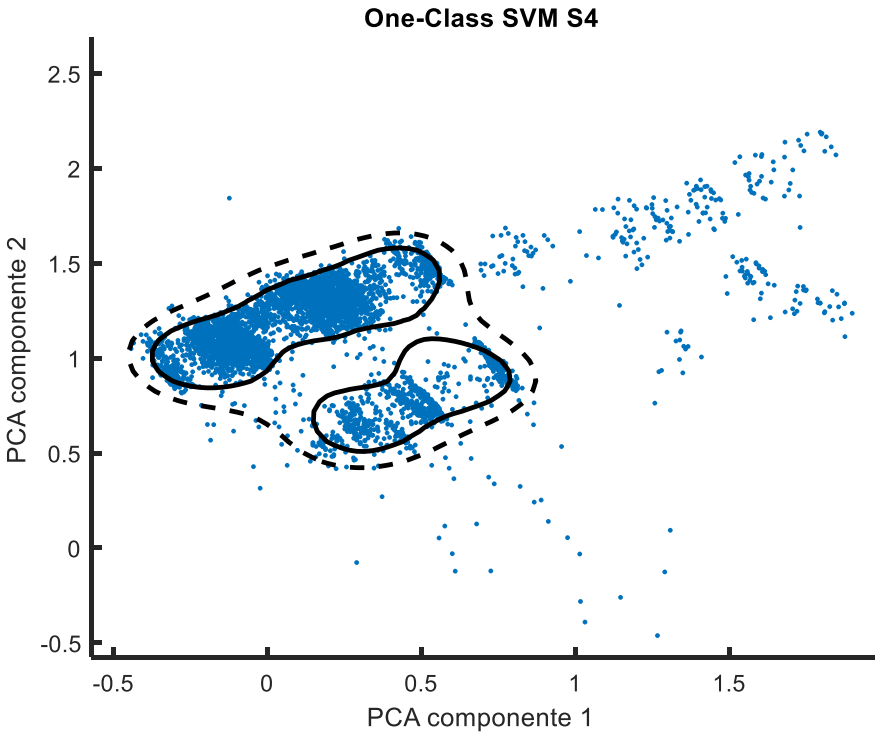

Ilustración 21. Límites de normalidad con One-Class SVM S4 ampliado

En esta sección se puede ver como existen dos modos de operación normal. El límite de normalidad configurado ha excluido el 13% de los datos.

3.3.1.5. Formación y enfriamiento

Por último, en la sección 5 se ha modelado con un parámetro *v* igual a 0.35, obteniendo el siguiente límite de normalidad.

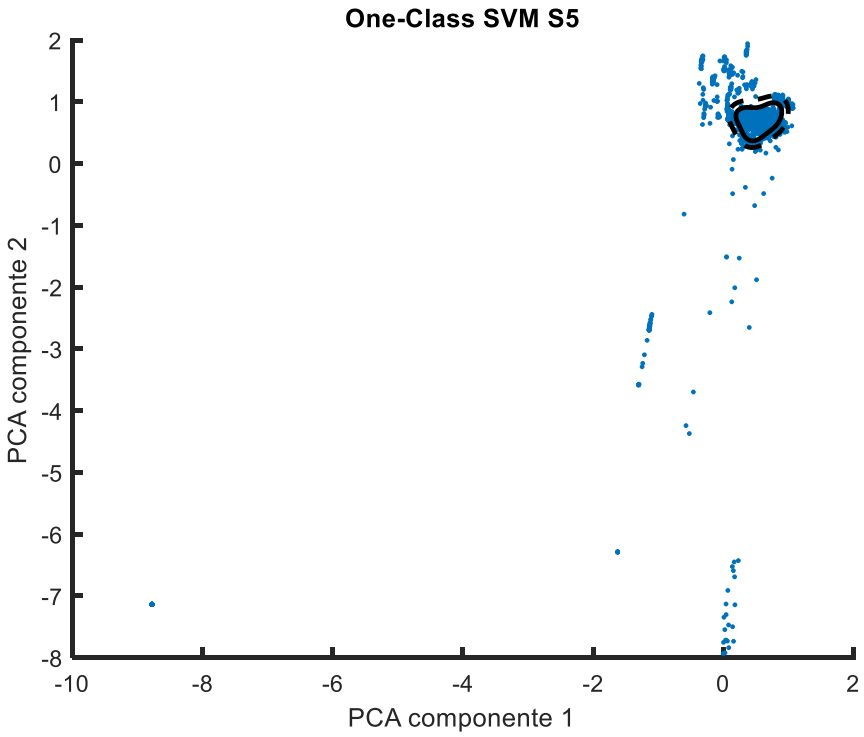

Ilustración 22. Límites de normalidad con One-Class SVM S5

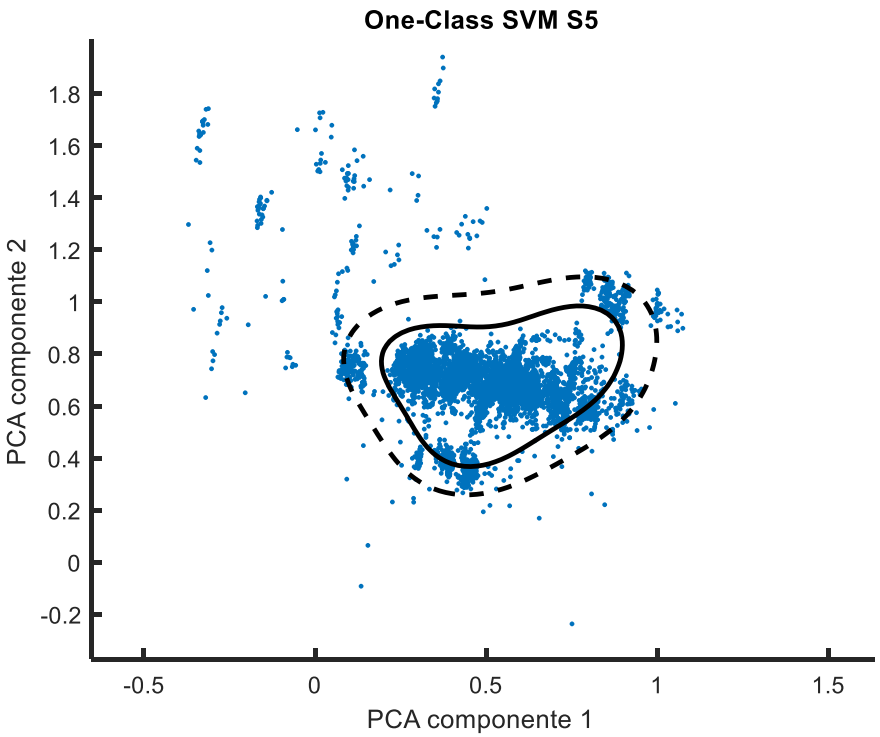

Ilustración 23. Límites de normalidad con One-Class SVM S5 ampliado

Este modelo del límite de normalidad ha descartado un total de 11% de los datos.

3.3.2. Índice de normalidad

Para esta sección se va a utilizar un conjunto de datos de palanquillas que tuvieron algún tipo de fallo y se va a evaluar por todo el modelo anterior para finalmente comprobar la cantidad de fallos que quedan fuera de estos límites. Este conjunto de datos tiene un tamaño de 95 palanquillas.

3.3.2.1. Horno de recalentamiento

En la sección 1 se han detectado 5 fallos fuera del límite de desviación, siendo un 5.3% de los datos, y 5 fallos entre el límite de normalidad y el límite de desviación, con un 5.3% del total de los datos. En la siguiente imagen podemos ver la distribución de los fallos en el plano.

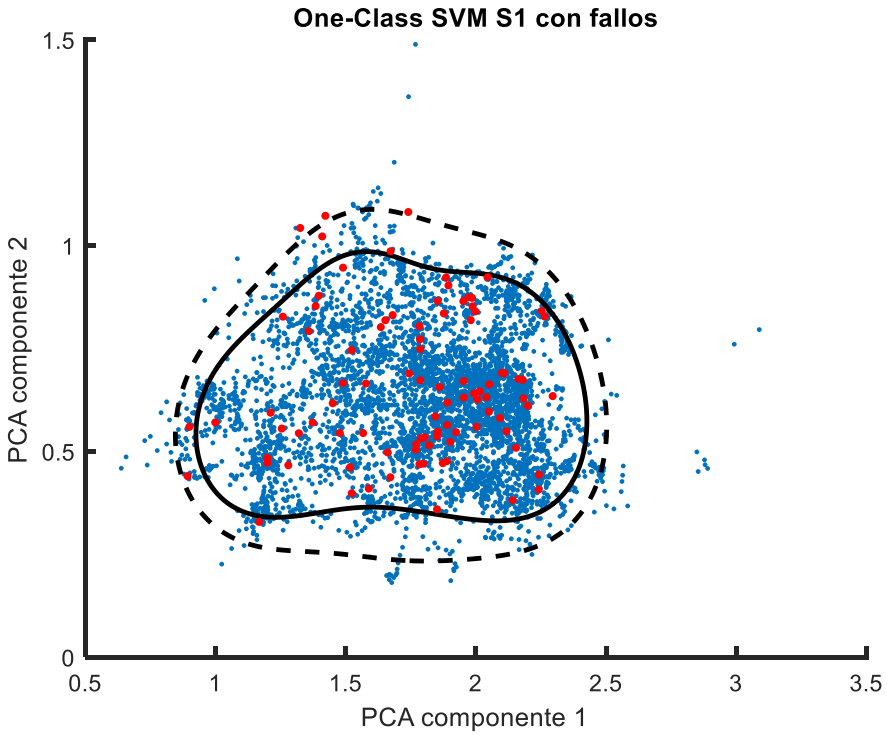

Ilustración 24. Límites de normalidad con fallos S1

3.3.2.2. Tren de desbaste

Después de analizar el conjunto de datos, para la sección dos, encontramos 8 palanquillas fuera del límite de desviación, conformando un 8.4% del total. Entre el límite de normalidad y el límite de desviación 4 palanquillas más, un 4.2% del total de los datos.

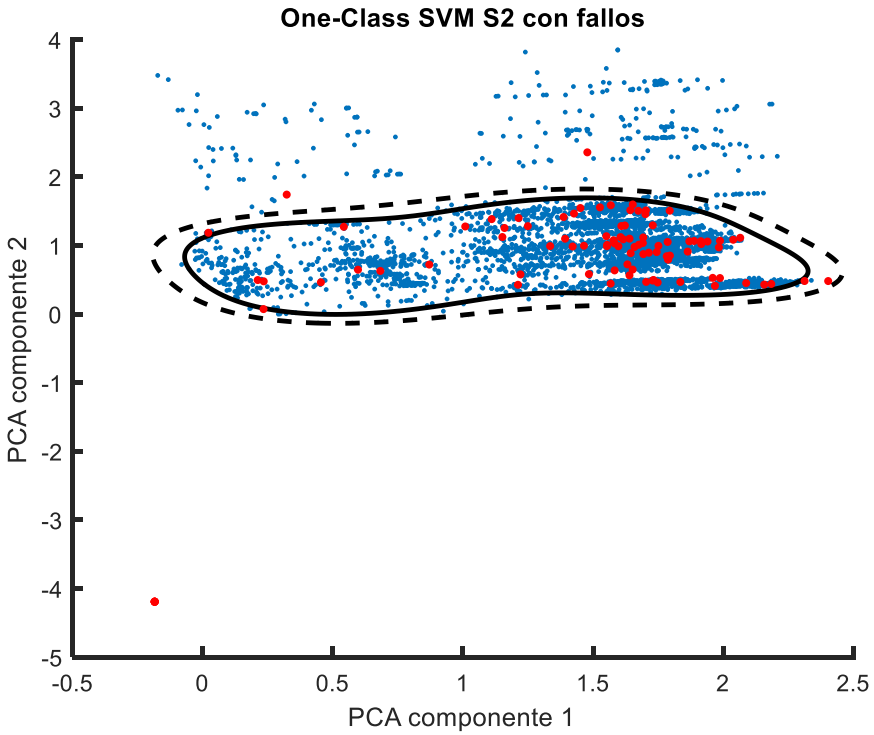

Ilustración 25. Límites de normalidad con fallos S2

3.3.2.3. Tren intermedio

Para la sección 3, se encuentran 8 fallos fuera del umbral de desviación, 8.4% del total. Entre el umbral de normalidad y el umbral de desviación se detectan 5 registros, 5.2% del total de los datos. En la siguiente ilustración se puede apreciar la distribución de los fallos en el plano consecuencia de la PCA.

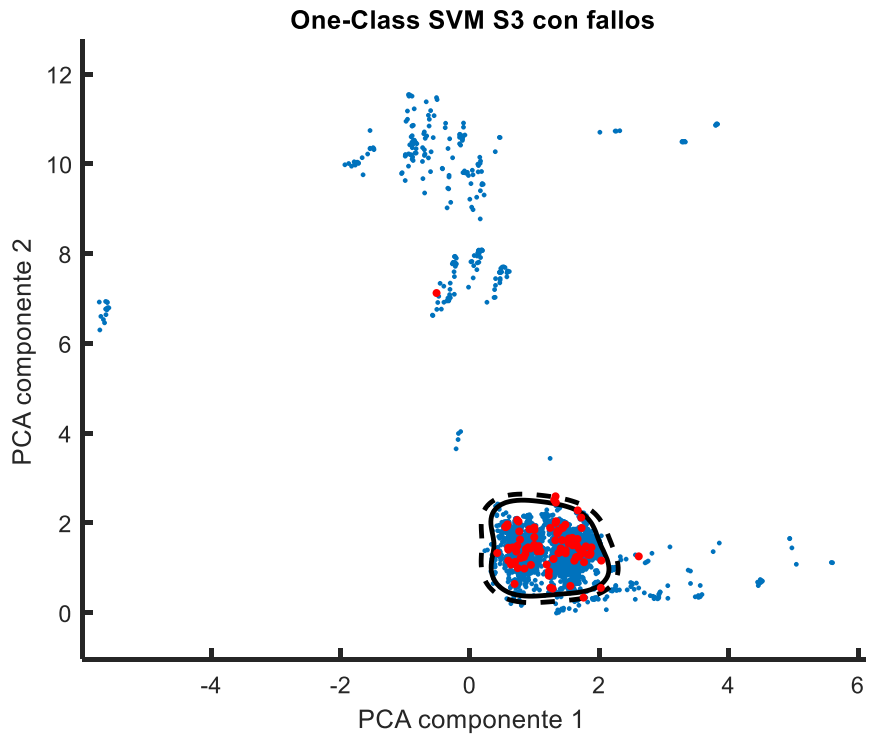

Ilustración 26. Límites de normalidad con fallos S3

3.3.2.4. Tren acabador

La sección 4 es algo más sensible a los fallos, fuera del umbral de desviación se ubican 35 fallos, un 36.8% del total de los fallos. Entre el umbral de desviación y el umbral de normalidad se localizan 24 fallos, un 25.3%. Seguidamente se muestra la disposición de los datos para esta sección.

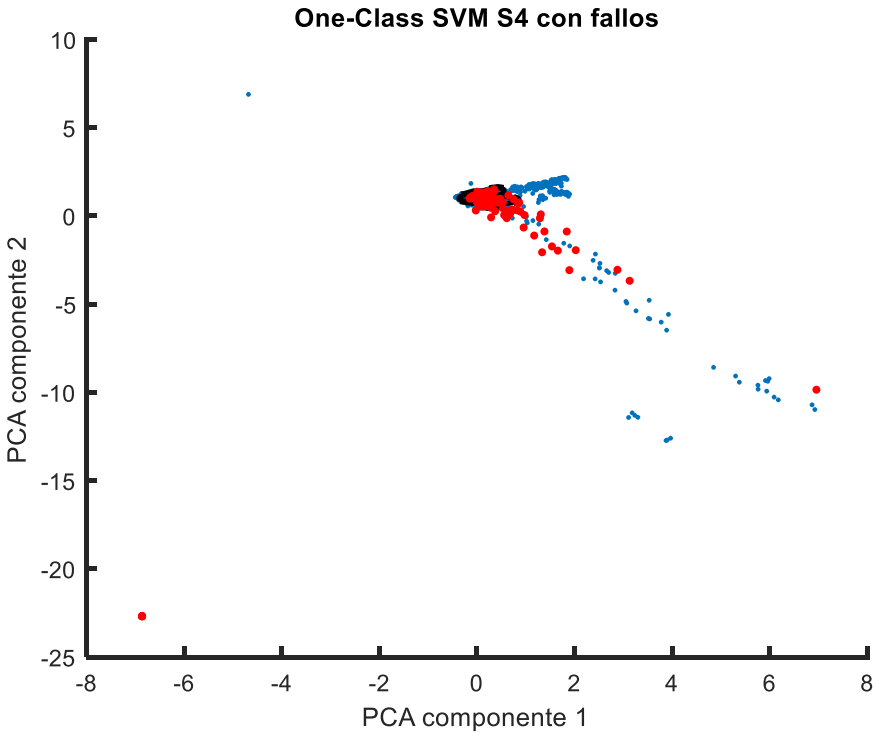

Ilustración 27. Límites de normalidad con fallos S4

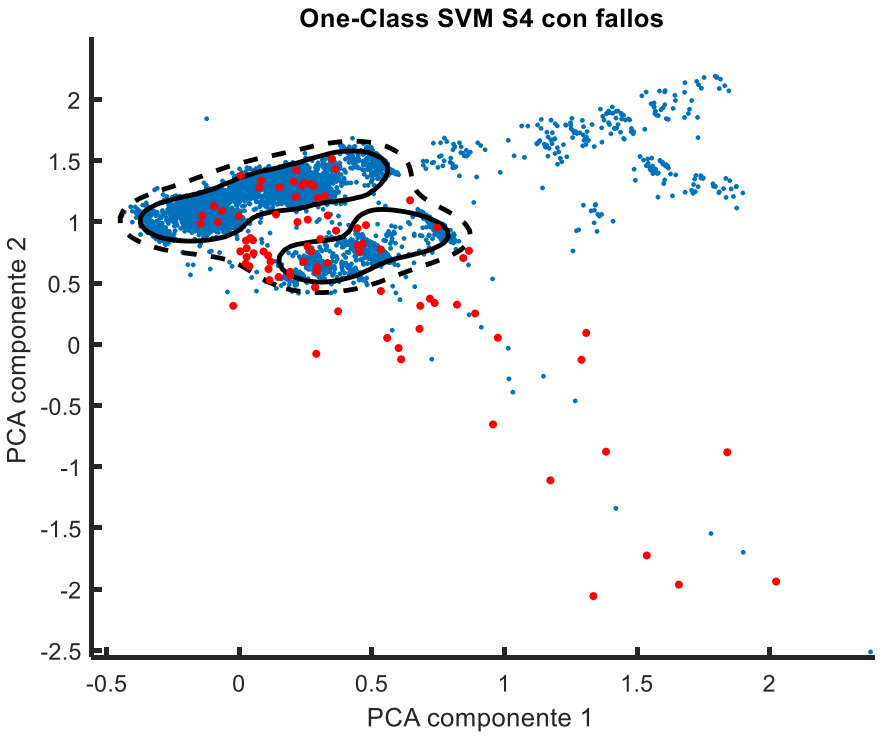

Ilustración 28. Límites de normalidad con fallos S4 ampliada

3.3.2.5. Formación y enfriamiento

La última sección, también sensible a los fallos, se detectan 25 muestras fuera del límite de desviación, 26.3% del total. Entre el límite de desviación y el límite de normalidad se detectan 13 registros, 37.7% del total de los fallos. En los siguientes gráficos se puede observar la disposición de las muestras.

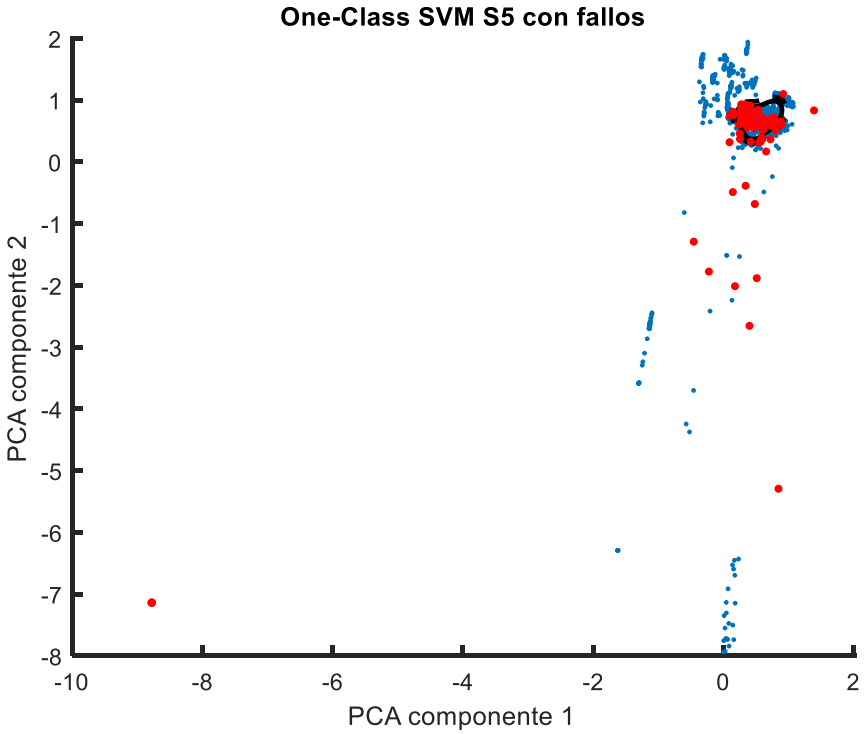

Ilustración 29. Límites de normalidad con fallos S5

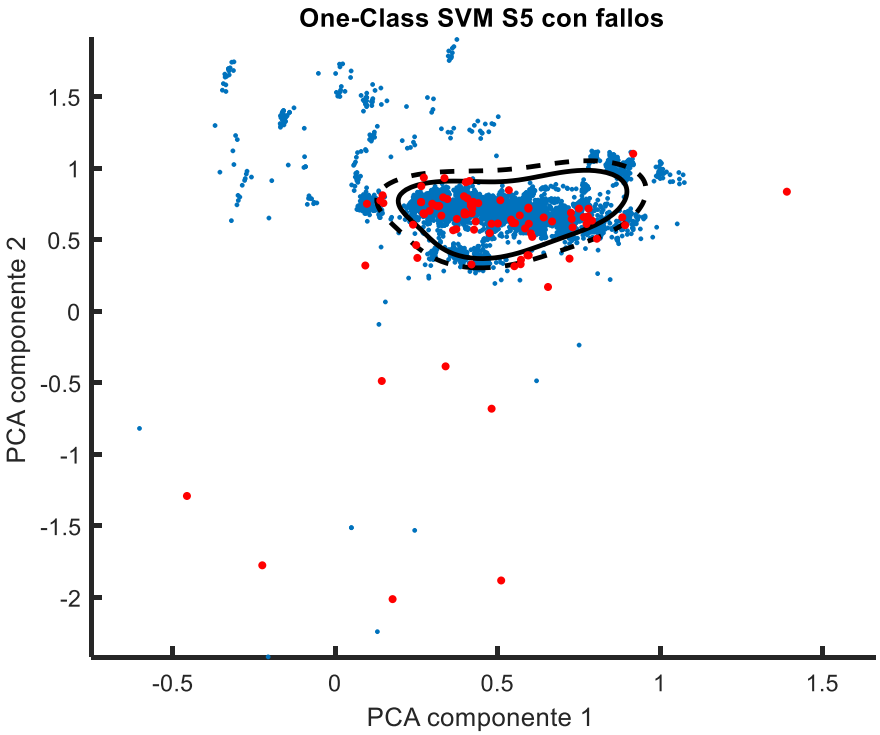

Ilustración 30. Límites de normalidad con fallos S5 ampliada

3.3.3. Índice de normalidad global

Una vez obtenido los índices de normalidad de cada sección por cada una de las muestras registradas se debe realizar el sumatorio de los índices de cada sección para así obtener el índice de normalidad global. Este índice nos permitirá detectar desviaciones de manera genérica del proceso.

En el siguiente gráfico se puede ver la diferencia de distribución de los índices de normalidad global entre los registros con palanquillas sin fallos y las muestras con fallos.

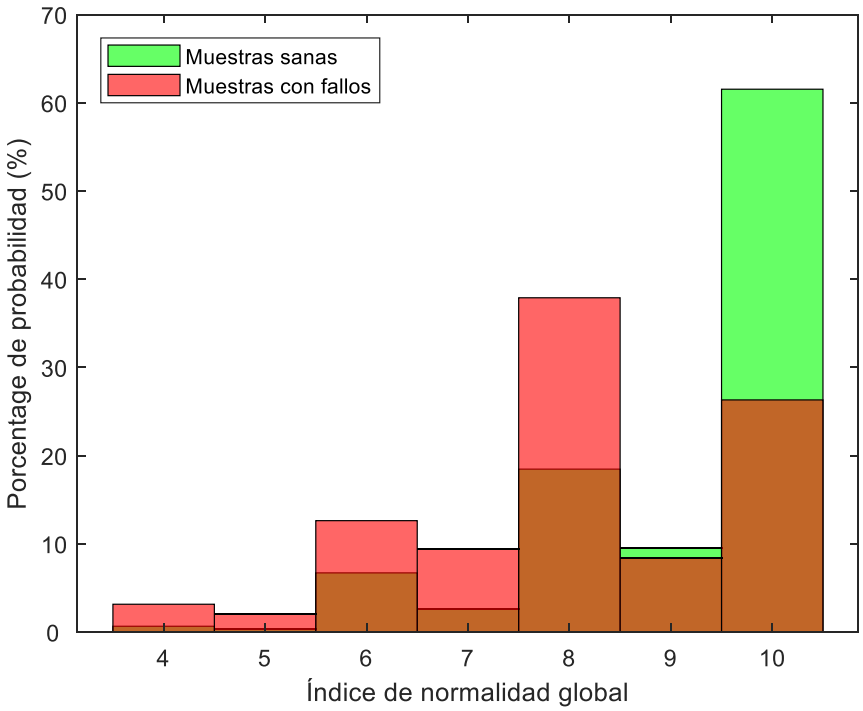

Ilustración 31. Distribución de probabilidad de palanquillas buenas contra fallos

Como se puede observar en el gráfico, la distribución de probabilidad de las muestras sanas se encuentra sobre los valores más grandes, el 60% de los datos obtienen un índice de normalidad global igual a 10, y el 70% de los datos se encuentra entre el 9 y el 10. Respecto a las muestras con fallos podemos observar cómo tienen una distribución que tiende a un índice de normalidad global más bajo. Encontrándose con un 68% de los datos inferiores o iguales a 8. El índice con mayor probabilidad de normalidad, respecto a las muestras con fallos, es el 8, con un 38%.

3.3.4. Visualización online

Suponiendo que la planta a la cual estamos analizando ya tiene implementada toda la solución y está introduciendo los datos a tiempo real en una BBDD a la cual poder acceder mediante Tableau. Para ello se propone la siguiente visualización, realizada con los datos utilizados en este estudio, para poder minorizar la cantidad de defectos en la producción.

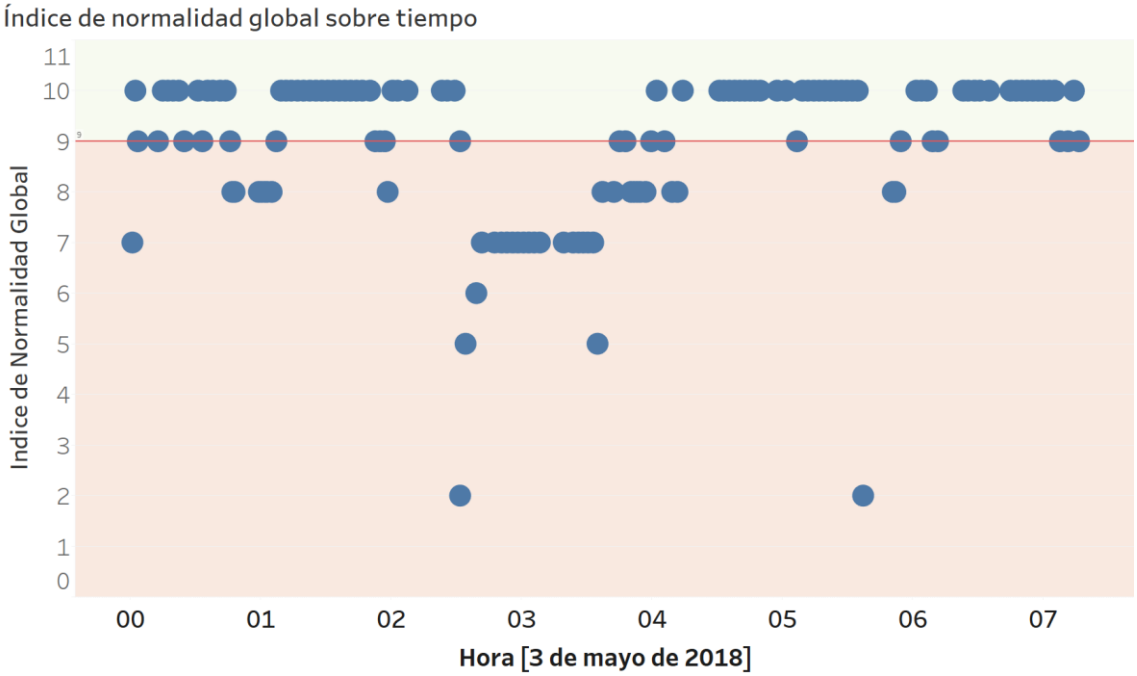

Ilustración 32. Propuesta visualización índice de normalidad global sobre el tiempo

En esta visualización se pueden ver el resultado del índice de normalidad global para cada una de las palanquillas sobre el eje tiempo. Se ha marcado un límite en sobre el índice de normalidad global igual a 9, dónde por debajo de ese índice se detecta un 70% de palanquillas con defectos y solo un 30% de palanquillas sin ningún tipo de defecto. Esto nos permitirá detectar fácilmente cierta desviación de la normalidad sobre el tiempo para así rectificar los parámetros de configuración o realizar mantenimientos. Pero para ello, se debe poder identificar cuáles son las secciones que están afectando a este índice de normalidad global. En el siguiente gráfico podemos ver la propuesta de visualización para poder identificar las secciones que están afectando.

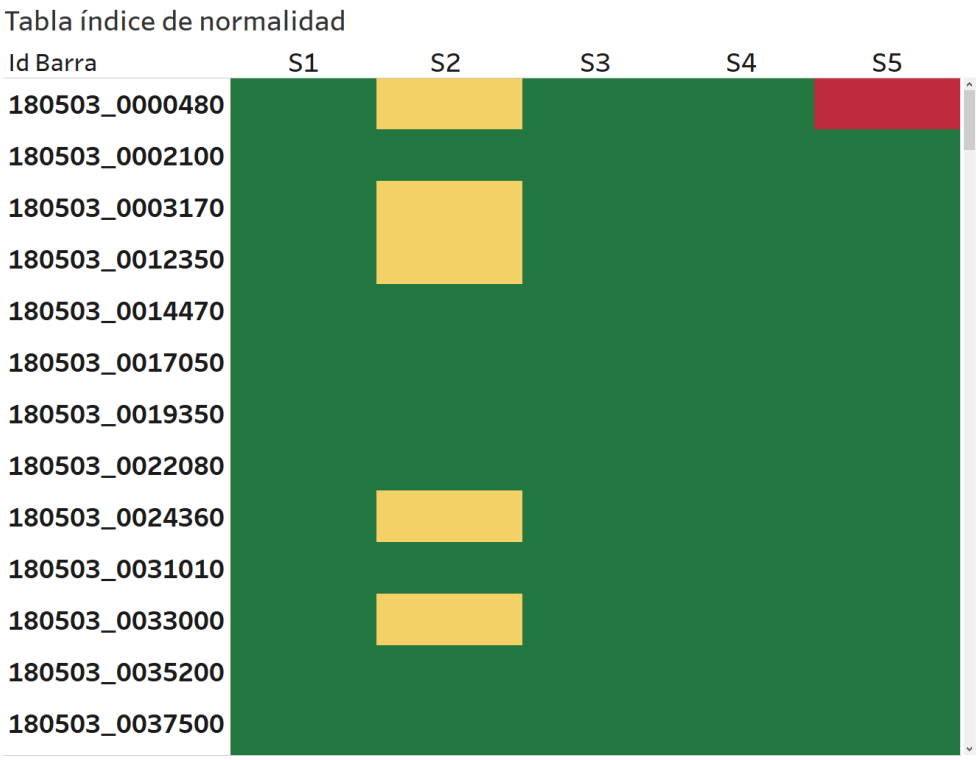

Ilustración 33. Propuesta de visualización de afectación al índice de normalidad global

En este segundo gráfico se puede observar la ID de la barra y con 3 colores el índice de normalidad por cada una de las secciones. El color verde indicara un índice igual a 2, eso significará que la palanquilla está dentro de la normalidad, en amarillo, un índice igual a 1, para representar aquellos valores que se están alejando de la normalidad, y finalmente, el rojo, aquellas que tengan un índice de normalidad igual a 0, representando todas aquellas palanquillas que quedan fuera de la normalidad.

Por último, se propone realizar un *dashboard* dónde intervengan los dos gráficos a la vez, de tal manera que si se selecciona alguno de los puntos de la primera visualización se van a filtrar los datos de la segunda así se pueden identificar fácilmente las palanquillas que se deseen. En la siguiente ilustración se puede observar el *dashboard* propuesto.

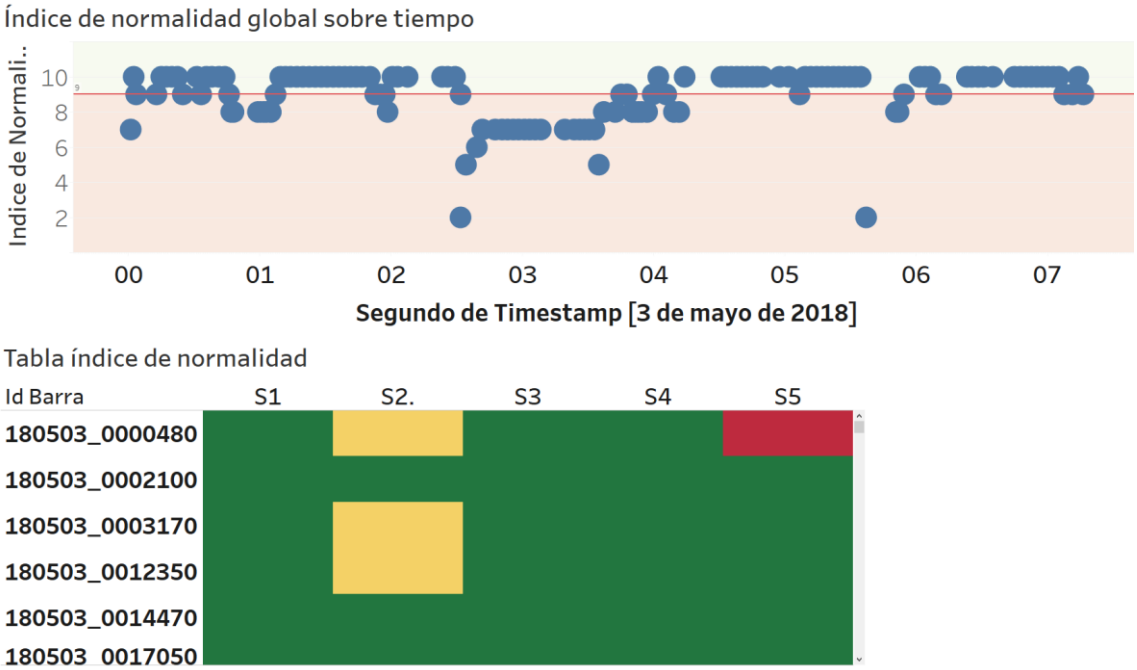

Ilustración 34. Propuesta de dashboard para la detección de errores

Las primeras barras que se observan en la tabla de índice de normalidad son las últimas introducidas en el proceso. En la siguiente ilustración se va a mostrar como quedaría si seleccionamos una de las zonas con un índice de normalidad global inferior a 9.

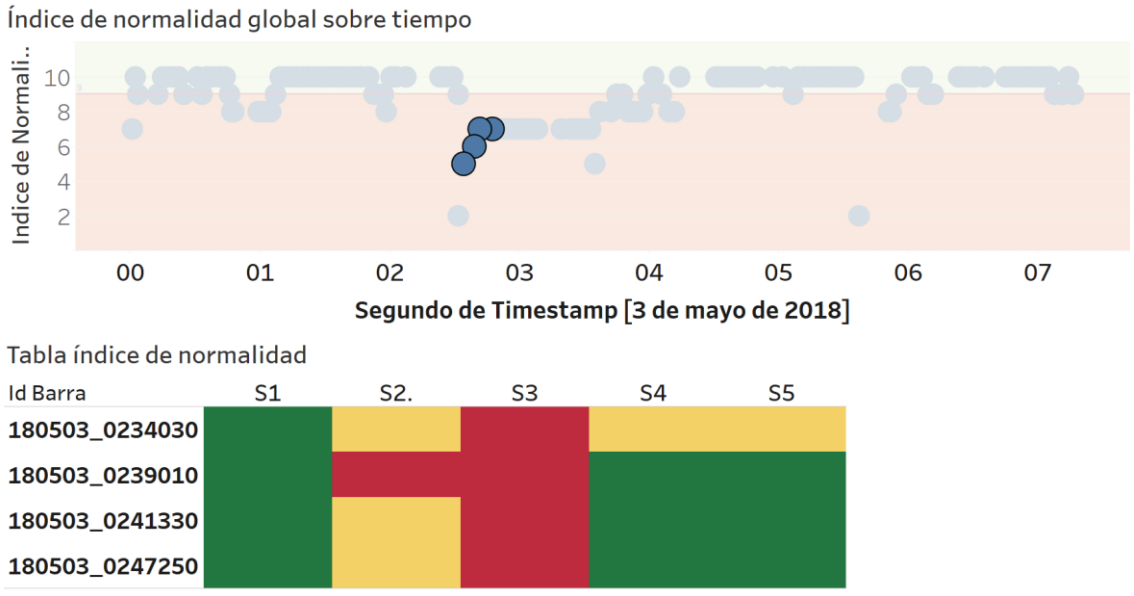

Ilustración 35. Propuesta de dashboard con selección de palanquillas

Como se puede observar se han seleccionado 4 palanquillas con un índice de normalidad bajo, fácilmente podemos detectar que la sección 3 está afectando muy negativamente al proceso, así como, la sección 2 también tiene cierta tendencia a estar algo desviada. Esto hace muy fácil la detección de anomalías en el proceso y las secciones implicadas a tiempo real.

4. RESUM DE RESULTATS

4.1. Resumen de presupuesto y viabilidad económica del estudio

El detalle del coste del estudio se puede encontrar en el presupuesto que se ha incluido en el trabajo, el coste total de este es de 16,800€. Para determinar la viabilidad económica del estudio, debemos estudiar si realmente es viable económicamente para la empresa. Debido a que la monitorización inteligente que se ha desarrollado en este estudio, se conseguiría disminuir la merma producida en 1%. Un 1% de la producción de 6 meses son 50 palanquillas, un total de 100 palanquillas anuales. Para calcular el coste ahorrado de debido a la aplicación de esta metodología se deberá calcular el coste de las 100 palanquillas.

El precio de un kg de acero es de aproximadamente 0,23€, considerando que cada palanquilla tiene un peso de 2,500 kg, cada palanquilla tiene un coste de materia prima de 575€, eso implica un ahorro anual de 57,500€, sin contabilizar el coste de la mano de obra. Con lo cual, en solo un año no solo se ha amortizado el coste del estudio, sino que se ha ahorrado 40,700€.

4.2. Implicaciones ambientales

Este estudio, salvo el coste energético del ordenador utilizado, no tiene ninguna implicación energética directa. También cabe remarcar que este estudio ayuda a reducir la merma de producción, con lo cual, se reduce el gasto de materia prima utilizada y el proceso es más eficiente.

4.3. Conclusiones

Después de realizar todo el conjunto de cálculos, visualizaciones y metodologías podemos concluir el estudio con varias observaciones sobre los mismos datos y los resultados de la aplicación de las metodologías pertinentes al estudio.

Sobre los datos objeto de análisis se han encontrado dos tipos de datos para tener en cuenta. En primer lugar, se ha encontrado cierta cantidad de datos con valores que no podrían

ser reales físicamente, lo que implica un mal registro del dato, la cantidad de malos registros se ha visto incrementada a partir de la sección 2. En segundo lugar, los *outliers*, o valores anómalos, aquellos valores que pueden llegar a influir a la hora de calcular los componentes de la PCA.

Respecto a los filtros utilizados para eliminar estos datos anómalos o defectuosos, se debe remarcar que de manera aproximada se ha eliminado a cada una de las secciones 1,000 muestras de un total de 5,587, descartando un 20% de los datos. Uno de los filtros aplicado no ha conseguido eliminar ninguna muestra. Concretamente el que seguía el siguiente criterio, todas aquellas muestras con la mitad de las medias de las señales de los diferentes sensores de una sección son igual a cero.

La metodología llamada *Principal Component Análysis* ha sido útil para la reducción de dimensionalidad y búsqueda de patrones. Al reducir la dimensionalidad, se pierde cierta cantidad de información, en este estudio el porcentaje medio de perdida es de alrededor del 55%.

En el estudio se ha fijado como límite de normalidad el 90% de los datos. Utilizando la técnica de *One-Class Support Vector Machine* y modificando el parámetro *v* en función de la distribución de los datos para cada una de las secciones se ha conseguido un límite medio del 12% de los datos.

Una vez resulto el modelo de normalidad sobre el espacio reducido con las componentes principales de la PCA se han analizado, sobre este modelo, 95 palanquillas con fallos. Respecto a este análisis podemos concluir que las secciones 4 y 5 son mucho más sensibles a los fallos, con un 60%, aproximadamente, de los fallos fuera de la normalidad, a diferencia de las otras 3 secciones.

Respecto al índice de normalidad global se ha observado que el 70% de las muestras sanas tienen un índice de 9 o 10 a diferencia del 30% de los datos con fallos. Con lo cual, podemos afirmar que los datos que tengan fallos van a tender hacia valores más bajos del índice de normalidad global.

También se ha ofrecido una propuesta visual, mediante Tableau, para una fácil detección de anomalías y la afectación de cada una de las secciones a tiempo real.

Por último, el coste de este estudio es rentabilizado en menos de un año, no solo eso, sino que consigue un ahorro de 40,700€ en el primer año. Cabe remarcar que este estudio no tiene ninguna implicación medioambiental negativa.

4.4. PROPUESTAS DE CONTINUACION DEL ESTUDIO

Como posibilidades para seguir estudiando la materia objeto de este estudio caben distintas posibilidades. Como primera opción se podría realizar un estudio de algoritmos para la predicción de desviaciones de tal manera que se conseguiría anticiparse a la desviación del proceso. Para ello se deberán incluir módulos de regresión para modelar a futuro variables criticas para anticipar el estado de producción a un horizonte de predicción de minutos.

Como segunda opción se añadirá un modulo de diagnostico de fallos en base a técnicas avanzadas de clasificación, como *Deep Neural Networks*, que complemente el modulo de predicción para saber que tipo de fallo está asociado a cada desviación, dependiendo del estado del proceso en las secciones.

5. BIBLIOGRAFÍA

[1] R. Davies, "Big data and data analytics The potential for innovation and growth," 2016.

[2] H. Lasi, P. Fettke, H.-G. Kemper, T. Feld, and M. Hoffmann, "Industry 4.0," *Bus. Inf. Syst. Eng.*, vol. 6, no. 4, pp. 239–242, Aug. 2014.

[3] "¿Qué es la Industria 4.0? | Deloitte España." [Online]. Available: https://www2.deloitte.com/es/es/pages/manufacturing/articles/que-es-la-industria-4.0.html. [Accessed: 04-Jun-2019].

[4] "Steel bars and wire rods from JFE."

[5] "Outline of Free Size Rolling for Hikari Wire Rod and Bar Mill."

[6] C. Gutierrez, "Partes del Tren de Alambron." [Online]. Available: https://es.scribd.com/document/372633359/Tren-Alambron. [Accessed: 04-Jun-2019].

[7] S. Theodoridis and K. Koutroumbas, *Pattern recognition*. Elsevier/Acad. Press, 2009.

[8] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, vol. 103. New York, NY: Springer New York, 2013.

[9] L. I. Smith, "A tutorial on Principal Components Analysis."

[10] J. Amat Rodrigo, "RPubs - Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE." [Online]. Available: https://rpubs.com/Joaquin_AR/287787. [Accessed: 04-Jun-2019].

[11] R. Vlasveld, "Introduction to one-class Support Vector Machines - Roemer's blog." [Online]. Available: http://rvlasveld.github.io/blog/2013/07/12/introduction-to-one-class-support-vector-machines/. [Accessed: 04-Jun-2019].

[12] Y. Chen, X. Zhou, T. S. Huang, N. Mathews Ave, and B. Institute, "ONE-CLASS SVM FOR LEARNING IN IMAGE RETRIEVAL."

[13] S. l. Russula, "Silat Wire Rod Mill." [Online]. Available: https://www.russula.com/es/news/news-silatwirerodmill-1.html. [Accessed: 07-Jun-2019].